Econometrics and Data Analysis for Accounting and Finance MN52080

Empirical Project 1 – Individual Work

Unit convenor: Qinye Lu
ql762@bath.ac.uk

DUE DATE: **Friday 12 December 2025, 14:00** (GMT)
Weight of Project Report to Overall Assessment – 40%

INSTRUCTIONS

- The whole report must not exceed 1,500 (±10%) words in length, excluding tables, graphics, references and appendices. Please state your 9-digit student number only at the beginning of the report.
- Your Word or PDF file of the final report must be submitted to Moodle. Save the file with the name of your 9-digit student number (e.g., 123456789.pdf). Your code, which can be a .py or .ipynb file, also needs to be submitted to Moodle with the name of your 9-digit student number followed by '_code' (e.g., 123456789_code.ipynb).
- The project criteria detailed below will be used to mark your final report submission. Please keep the criteria in mind when preparing your report.
- Students may apply for a coursework extension. Detailed academic regulations can be found via the below link: https://www.bath.ac.uk/guides/coursework-extensions/#how-to-apply-for-a-coursework-extension

PURPOSE OF THIS EMPIRICAL PROJECT

One of the central questions in finance is how financial and economic variables explain the movement of stock market returns or the equity premium, as explored in highly influential studies such as Goyal and Welch (2008) and Goyal, Welch and Zafirov (2024). Researchers have investigated various predictive factors, including dividend yields, book-to-market ratio, interest rates, and macroeconomic indicators, to understand and forecast these movements (Campbell & Thompson, 2008; Cochrane, 2011; Fama & French, 1989). This question remains pivotal for both academic research and practical investment strategies, given the implications for asset pricing and portfolio management.

In recent years, climate vulnerability has emerged as a critical macro-financial variable that may influence asset returns and risk premia. Climate vulnerability reflects the degree to which a country's economy and infrastructure are exposed to, and unable to cope with, the adverse effects of climate change - such as extreme weather events, resource scarcity, and temperature shocks. It captures primarily physical risks arising from climate hazards, including floods, storms, and droughts, and the economic sensitivity to such events (Batten, Sowerbutts & Tanaka, 2020). A growing body of evidence suggests that higher climate vulnerability can increase economic volatility, reduce productivity, and elevate borrowing costs, particularly in developing and disaster-prone economies (Kling et al., 2025; Torres, 2024). Moreover, climate-related shocks and news have been shown to influence asset prices and risk premia, as financial markets increasingly price in climate risks (Engle et al., 2020). These risks may translate into higher risk premia and lower equity valuations as investors demand compensation for heightened uncertainty related to climate exposure.

This project examines how climate vulnerability - measured by the Vulnerability Score of the Notre Dame Global Adaptation Initiative (ND-GAIN) Index - along with other traditional factors, affects stock market returns and how this impact varies during periods of heightened economic risk. The objective is to assess whether increases in the ND-GAIN Vulnerability Score have a significant impact on stock market returns and to analyse the robustness of this relationship over time.

You are expected to apply the regression methods learned in the course to estimate a regression equation for the excess stock returns of the US market, addressing potential econometric issues such as heteroskedasticity, autocorrelation, and multicollinearity. Additionally, you should provide economic reasoning to support your findings and demonstrate proficiency in statistical analysis using Python (in Jupyter Notebook or Spyder).

DATA

The Notre Dame Global Adaptation Initiative (ND-GAIN) Country Index data for the Vulnerability scores is sourced from the official ND-GAIN website [1]. This dataset is provided in the file *vulnerbility.csv* and contains global annual data. The ND-GAIN Vulnerability Score ranges from 0 (low vulnerability) to 1 (high vulnerability) and captures a country's sensitivity and exposure to climate hazards across sectors such as food, water, health, ecosystem services, human habitat, and infrastructure. For example, a score of 0.27 indicates relatively low vulnerability, suggesting strong adaptive capacity and limited exposure to climate risks; a score of 0.49 reflects moderate vulnerability, implying a noticeable degree of climate-related risk but some resilience; whereas a score of 0.65 denotes high vulnerability, meaning the country is more exposed and less able to cope effectively with adverse climate impacts.

[1] Source: https://gain.nd.edu/our-work/country-index/download-data/

For this project, you should extract the United States data only, as the analysis focuses on the US market. If your analysis uses quarterly or monthly data, you are expected to extend the annual ND-GAIN series accordingly. For instance, if the annual ND-GAIN Vulnerability Score for 20XX is 0.49, you should assign the same value (0.49) to all months or quarters within 20XX. Other data are available in *Goyal_Welch_data.xlsx*. Definitions of these variables, along with their abbreviations as they appear in *Goyal_Welch_data.xlsx*, are listed below, and you can read the original paper for further details. You will need to refer to the study of Welch and Goyal (2008) and use the provided data to properly create additional variables used in your regression model, such as the Dividend Price Ratio (*d/p*), Dividend Yield (*d/y*), Dividend Payout Ratio (*d/e*), Default Yield Spread (*dfy*), Term Spread (*tms*), and Default Return Spread (*dfr*).

- **Price** (*Index*) is the S&P 500 index price.
- **Dividends** (*D12*) are 12-month moving sums of dividends paid on the S&P 500 index. The data are from Robert Shiller's website from 1871 to 1987. Dividends from 1988 and after are from the S&P Corporation.
- **Earnings** (*E12*) are 12-month moving sums of earnings on the S&P 500 index.
- **Book-to-Market Ratio** (*b/m*) is the ratio of book value to market value for the Dow Jones Industrial Average.
- **Treasury Bills** (*tbl*) rates from 1920 to 1933 are the *U.S. Yields On Short-Term United States Securities, Three-Six Month Treasury Notes and Certificates, Three Month Treasury* series in the NBER Macrohistory data base. Treasury-bill rates from 1934 onwards are the *3-Month Treasury Bill: Secondary Market Rate* from the economic research data base at the Federal Reserve Bank at St. Louis (FRED).
- **Corporate Bond Yields** on AAA and BAA-rated bonds (*AAA* and *BAA*): They are collected from Federal Reserve Economic Data (FRED).
- **Long Term Yield** (*lty*): Long-term government bond yields for the period 1919 to 1925 is the *U.S. Yield On Long-Term United States Bonds* series from NBER's Macrohistory database. Yields from 1926 and after are from Ibbotson's *Stocks, Bonds, Bills and Inflation Yearbook*.
- **Net Equity Expansion** (*ntis*) is the ratio of twelve-month moving sums of net issues by NYSE listed stocks divided by the total market capitalization of NYSE stocks.
- **Risk-free Rate** (*Rfree*): The risk-free rate is the Treasury-bill rate from 1920 onwards. Because there was no risk-free short-term debt prior to the 1920s, they were estimated based on the Commercial paper rates.
- **Inflation** (*infl*): Inflation is the Consumer Price Index (All Urban Consumers) from the Bureau of Labor Statistics.
- **Long Term Rate of Return** (*ltr*): Long-term government bond returns are from Ibbotson's Stocks, Bonds, Bills and Inflation Yearbook
- **Long-term corporate bond** (*corpr*): Long-term corporate bond returns are from Ibbotson's Stocks, Bonds, Bills and Inflation Yearbook.
- **Stock Returns:** The total rate of return on the US stock market is the S&P 500 index returns from the Center for Research in Security Press (CRSP). Stock returns are the continuously compounded returns on the S&P 500 index, including dividends. In the provided data file, both the returns on the S&P 500 index, including dividends (*CRSP_SPvw*) and excluding dividends (*CRSP_SPvwx*) are provided.

REQUIREMENTS

- You are required to develop a *multiple* regression model to explain the excess returns of the US market. Define your research question or the objectives of your study and explain the relationships and expected signs of variables using sound economic or financial reasoning. You must include at least three independent variables, one of which should be the ND-GAIN Vulnerability Score, and the others selected or constructed based on insights from relevant literature. Present your econometric model clearly and justify your variable choices. (Please note, we do not require out-of-sample regression in this project.)

- You should prepare a suitably large monthly or quarterly dataset for your empirical study. For example, a dataset with 360 observations spanning 30 years of monthly data from January 1991 to December 2020 would be appropriate. However, you are allowed to choose a different time period. Additionally, you need to properly integrate the ND-GAIN Vulnerability Score with the rest of your dataset. When selecting variables, check for missing values and clearly explain how you address any data gaps. You should also consider how to handle potential extreme or outlier values to ensure the robustness of your results[2]. Define the type of data you are using (e.g., pooled cross-sectional data), and discuss any limitations associated with using economic or financial data. You do not need to provide detailed definitions for the variables given in the provided Excel files as they are already explained above but make sure to refer to variable names correctly and clearly explain any new variables you construct.

Regression Analysis and Diagnostic Tests

- Use the appropriate test to identify the presence of multicollinearity within the data. Interpret the results and try to resolve it if problematic multicollinearity is found.
- Perform a multiple regression analysis to determine the values for the parameters. Explain the significance of the parameters in the context of 'ceteris paribus' and the 'partial effect.' Interpret the test statistics for parameters and explain their implications based on your results. Discuss the potential consequences of omitting a key explanatory variable.
- Explain the meaning of R^2 and adjusted R^2 concerning your sample of data.
- Interpret the F-statistic of your regression, explaining the suitability of your model.
- Determine whether the error terms are normally distributed.
- Determine whether heteroskedasticity is present in the errors using the appropriate tests. Interpret the results of each test, discuss the test advantages and limitations, and suggest actions to take if heteroskedasticity is detected.
- Determine the presence of autocorrelation in the errors by conducting appropriate tests. Interpret your results of each test, discuss the advantages and limitations of the test, and suggest actions to take if autocorrelation is detected.
- The economic conditions also play an important role in explaining excess market returns. Create a dummy variable to indicate whether it is a recession period and run your regression with a proper interpretation. You do not need to re-run the previous diagnostic tests when introducing this dummy variable. A list of US recessions can be found on The Wikipedia (https://en.wikipedia.org/wiki/List_of_recessions_in_the_United_States). Additionally, consider whether a structure break may exist and provide evidence to support your conclusion.

[2] Hint: you can winsorize to limit extreme values to 1% or 5%.

MARKING SCHEME

Item	Marks
1. Outline the question of interest.	5
2. Explain the economic or financial reasoning behind your model and present the econometric specification.	12
3. Data formation, description and omitted data	12
4. Regression Analysis (incl. parameter estimation, R2, t-test for each parameter and F-statistic etc)	15
5. Diagnostic Tests (incl. normal distribution, multicollinearity, heteroskedasticity, autocorrelation)	32
6. Dummy variable and structure break	18
7. Presentation, Reference, Format	6
TOTAL	**100**

REFERENCE

Batten, S., Sowerbutts, R. and Tanaka, M., 2020. Climate change: Macroeconomic impact and implications for monetary policy. *Ecological, societal, and technological risks and the financial sector*, pp.13-38.

Campbell, J.Y. & Thompson, S.B., 2008. Predicting excess stock returns out of sample: Can anything beat the historical average? *Review of Financial Studies*, 21, pp.1509-1531.

Cochrane, J.H., 2011. Presidential address: Discount rates. *Journal of Finance*, 66, pp.1047-1108.

Engle, R.F., Giglio, S., Kelly, B., Lee, H. and Stroebel, J., 2020. Hedging climate change news. *The Review of Financial Studies*, *33*(3), pp.1184-1216.

Fama, E.F. & French, K.R., 1989. Business conditions and expected returns on stocks and bonds. *Journal of Financial Economics*, 25, pp.23-49.

Goyal, A. & Welch, I., 2008. A comprehensive look at the empirical performance of equity premium prediction. *Review of Financial Studies*, 21, pp.1455-1508.

Goyal, A., Welch, I. and Zafirov, A., 2024. A comprehensive 2022 look at the empirical performance of equity premium prediction. *The Review of Financial Studies*, *37*(11), pp.3490-3557.

Kling, G., Lo, Y.C., Murinde, V. and Volz, U., 2025. Climate vulnerability and the cost of debt. *Oxford Open Economics*, *4(1)*, pp.1-14.

Torres, M.F., 2024. Brazil's vulnerability to climate change: an analysis based on the University of Notre Dame's Global Adaptation Initiative (ND-GAIN). *Conjuntura internacional*, *21*(1), pp.12-20.