

Replication Package for: Recovering Investor Expectations from Demand for Index Funds

Mark Egan
Harvard University
and NBER[‡]

Alexander MacKay
Harvard University^{*}

Hanbin Yang
Harvard University[†]

October 7, 2021

We combine several publicly available databases and non-public data we accessed through Wharton Research Data Service (WRDS) and Bloomberg under our academic licenses. We provide copies of public data under their respective folder in /Data/ when possible. The user agreements for many of the following data do not allow us to publish the underlying data, in which case we provide reference and url link where researchers can download the data. Researchers can request WRDS and Bloomberg accounts through academic licenses with their institutions.

Stata and R codes to replicate our results are provided in /Code/. Running 00_master.do will execute all scripts in order by the two-digit prefix. Outputs are generated and saved in /Output/Figure/, /Output/csv/, and /Output/Tex/.

1 Data Availability Statement

1.1 Public Data

We assembled the following data from public sources. When permitted, we include the raw data in /Data/. The use agreements for many of the following data do not allow us to publish or broadcast the underlying data. We provide the citation and url to the sources, where researchers should be able to download these data. Citations appear in the References section at the end of this README.

^{*}Harvard University, Harvard Business School. Email: amackay@hbs.edu.

[†]Harvard University, Harvard Business School and Department of Economics. Email: hyang1@g.harvard.edu.

[‡]Harvard University, Harvard Business School. Email: megan@hbs.edu.

- Yahoo Finance (2020): S&P 500 index and VIX Index. Raw data for S&P 500 index provided in /Data/YahooFin/GSPC.csv (2005-2019), /Data/YahooFin/SP500_2020.xlsx (2020). Raw data for VIX provided in /Data/YahooFin/VIX.csv (2005-2019) and /Data/YahooFin/VIX_2020.xlsx (2020).
- Board of Governors of the Federal Reserve System (2020): Retail money market mutual funds. A copy of the data is provided in /Data/Fed/RMFSL.csv and /Data/Fed/WRMFSL.csv.
- Dong et al. (2020): COVID cases and deaths from the COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University. A copy of the data is provided in /Data/covid/us_death_johnshopkins.xlsx.
- Kenneth R. French - Data Library (2019): Market excess return.
- Survey of Consumers (2019): University of Michigan Survey of Consumers data.
- AAI (2019): American Association of Individual Investors (AAII) Investor Sentiment Survey data.
- United States Stock Market Confidence Indices (2019): Originally collected by Robert Shiller and now maintained by Yale School of Management.
- Shiller (1992): Stock price, dividend, and earning series retrieved from Robert Shiller website (<http://www.econ.yale.edu/~shiller/data.htm>).
- Lettau and Ludvigson (2001): *cay* data, retrieved through Martin Lettau's personal website (<https://sites.google.com/view/martinlettau/data>).
- Wells Fargo/Gallup (2019): Investor and retirement optimism index.
- Survey of Professional Forecasters (2019): Forecast data used in our analysis is provided in /Data/SurveyProfessionalForecast/.
- Baker et al. (2019): US equity market volatility index, retrieved from https://www.policyuncertainty.com/EMV_monthly.html.
- Ludvigson et al. (2020): Financial uncertainty index, retrieved through Sydney Ludvigson's website <https://www.sydneyludvigson.com/macro-and-financial-uncertainty-indexes>.

1.2 Non-Public Data

We use the following data through our academic license with WRDS and Bloomberg. These are commonly used research databases. Researchers can request WRDS and Bloomberg accounts through their institutions.

- ETF Global (2020): ETF descriptions, total net assets, and fees. Researchers can download ETF global data to folder /Data/ETFGlobal. Please refer to 01_data_load.do for further information.
- CRSP Mutual Fund (2020): ETF total net assets and fees. Researchers can download CRSP Mutual Fund global data to folder /Data/CRSP. Please refer to 01_data_load.do for further information.
- OptionMetrics (2020): Option prices to replicated Martin (2017), and realized volatility indices. Researchers can download option metrics data to folder /Data/SVIX2. Please refer to 01_data_load.do and 17_option_metrics.do for further information.
- Bloomberg L.P. (2020): We retrieved data on ETF net asset value, total net asset, trading volume, shares outstanding, institution ownership, and volatility indices for asset classes other than S&P 500, and GDP growth through Bloomberg Terminal. Please refer to 01_data_load.do, 04_data_load_daily.do, and 18_expectation_analysis_construct_data.do for further information. Researchers can save respective data in /Data/Bloomberg/.

We also obtain Duke CFO Survey data (Graham and Harvey (2011)). We are not allowed to publish or broadcast the underlying data. Researchers can contact John Graham and Campbell Harvey at Duke CFO Global Business Outlook. For further information on how we use this data, please refer to 08_expectation_analysis_construct_data.do.

2 Computation Requirements

2.1 Software

We use Stata version 17 and R version 4.0.5

R code requires the following packages

- Matrix(v1.3.2)
- numDeriv(v2016.8.1.1)
- skewt(v0.1)
- latex2exp(v0.5.0)
- openxlsx(v4.2.3)
- tidyr(v1.1.3)
- XML(v3.99.0.6)
- haven(v2.3.1)

- randtoolbox(v1.30.1)
- robustbase(v0.93.7)
- Rcpp(v1.0.7.2)
- RcppArmadillo(v0.10.2.2.0)
- foreign(v0.8.81)

2.2 Memory and Runtime

The code was run on a MacBook Pro with 4 Cores and 16GB of memory. The entire process takes roughly 5 days. 10_etf_demand_est_sml.R and using 07_etf_demand_est.R for model number 11 consume the majority of runtime. The other do files and R codes are not computationally intensive. Please refer to section Description of Code and Replication Instruction for further information

3 Description of Code and Replication Instructions

3.1 Instructions for Replicators

Running 00_master.do will execute all scripts in order by the two-digit prefix. Please see the comments in 00_master.do for a list of stata and R codes. Before running, it is necessary to set the global variables \$BASE and \$R_PATH to the project directory and the system path for R in 00_master.do. See the script for more details.

3.2 Description of Code

01_data_load.do: Aggregate ETF global, CRSP, Bloomberg, some additional macro times series to construct market shares and fees for ETFs tracking S&P 500 as well as other markets at monthly level from 2008 to 2018.

Process raw data in the following folders:

- /Data/ETF_List
- /Data/ETFGlobal
- /Data/CRSP
- /Data/Bloomberg
- /Data/YahooFin
- /Data/Fed

02_data_estimation.do: Aggregate ETF level market shares and fees to leverage category level, and prepare data for estimation in R.

03_summary_statistics.do: Plot total AUM and trading volume of S&P 500 ETF overtime. Produce summary statistics at leverage category level. Plot time series of market share and fees across leverage category over time.

Produce the following exhibits:

- Table 1
- Figure 1, 2, A11

04_data_load_daily.do: Aggregate Bloomberg data to construct market shares for ETFs tracking S&P 500 at daily level for 2020. Aggregate to leverage category level and prepare data for estimation in R.

Process raw data in the following folders:

- /Data/Bloomberg
- /Data/YahooFin
- /Data/covid
- /Data/Fed

05_config.R: Define which market and model specification for estimation. Set up position of parameters in inner vs outer loop optimization, initial conditions, and a tag for output version. This script should be loaded before running any other R code.

In this script, we load model specifications from `model_spec.csv` and initial values for each parameter in `param_init.csv`.

06_programs.R: R functions for estimation and generating plots.

07_etf_demand_est.R: Main wrapper to perform estimation using maximum likelihood for monthly level data. This script saves estimation in a R file that can be loaded later on to generate outputs.

08_compute_se.R: Load the estimation from `07_etf_demand_est.R` to compute standard errors. We only compute standard errors for two baseline specifications.

09_sml_programs.R: R functions for estimation using simulated maximum likelihood.

10_etf_demand_est_sml.R: Perform estimation using simulated maximum likelihood. This is for specifications that allow heterogeneities in both expectation and risk aversion. This script saves

estimation in a R file that can be loaded later on to generate outputs.

11_ett_demand_est_daily.R: Perform estimation using maximum likelihood for daily level data. Produce the following exhibits:

- Figure 9

12_programs_outputs.R: Additional R functions to wrap around functions in 06_programs.R which aggregate estimated parameters and produce desired outputs for each specification.

13_generate_outputs.R: Generate csv and pdf outputs for each market and model specification. This produces main figures for each specification: the distribution of recovered expectations, measure of fit comparing log shares in the data and predicted, time series of risk aversion as well as dispersion for specifications allow for time-variation. For our main specifications, output figures also include the pdf plot providing intuition for distribution of expectation, time series of estimated belief parameter as well as standard errors, and CDF plot showing identifications

Produce the following exhibits:

- Table 1
- Figure 3,4,5,6,8, A1, A3, A4, A5, A6, A7, A8, A9, A10

14_outputs_across_specs.R: Produce several csv file for latex tables and numbers discussed in the manuscript: Comparing goodness of fit (including AIC, BIC), estimated parameters, distribution of expectation, elasticities, etc across multiple specifications of the S&P 500 market. Comparing parameter estimates and standard errors in two main specifications of the S&P 500 market. Correlation of mean beliefs and risk aversion across multiple specifications of the S&P 500 market. Comparison between VIX and realized vol, between baseline and combining 2/3x leverages.

Produce the following exhibits:

- Table 3, A1

15_aggregate_r_output.do: Take csv outputs from each market generated in 13_generate_outputs.R and aggregate into one dta file.

16_correlation_asset_classes.do: Take the aggregated dta from 15_aggregate_r_output.do to compute correlation of mean beliefs across markets.

Produce the following exhibits:

- Table A2

17_option_metrics.do: Load option data from Option Metrics in /Data/SVIX2 to replicate the lower bound of equity premium from Martin (2017) and compare with estimated belief distribution.

Produce the following exhibits:

- Figure A2

18_expectation_analysis_construct_data.do: Consolidate survey data and expected return data from multiple sources:

- Duke CFO Global Business Outlook: /Data/CFOSurvey/
- Gallup Optimism Index: /Data/Gallup/
- University of Michigan Survey of Consumers: /Data/MichiganSurvey/
- AAI Sentiment Survey: /Data/AAI/
- Shiller Individual Confidence Index: /Data/Shiller/
- Survey of Professional Forecasts: /Data/SurveyProfessionalForecast/
- Baker Bloom Davis US equity market volatility index: /Data/BakerBloomDavis/
- Ludvigson, Ma, and Ng (2020) Financial uncertainty index:
/Data/MacroFinanceUncertainty_202008_update
- Market excess return from Fama-French library: /Data/FF/
- Model based expected return, dividend to price ratio from /Data/Shiller/ and cay from /Data/CAY/
- Macro variables including volatility index and GDP growth from Bloomberg: /Data/Bloomberg

19_expectations_analysis.do: Load parameter estimates and distribution of expectation from 15_aggregate_r_output.do, survey data well as historical and future market returns from 18_expectation_analysis_construct_data.do.

Compare the recovered expectation vs survey, historical return, future return, and model based expected return. Estimate regression coefficients showing the correlation of our estimates with survey, historical return, etc, and also produce binned scatter plots.

Also estimate regressions to understand the extrapolative pattern of expectation and the implication on heterogeneity.

Produce the following exhibits:

- Table 4, 5, 6, 7, 8, 9, 10, A3, A4, A5, A6, A7, A8
- Figure A12, A13

20_impulse_rsp.R: Take parameter estimates overtime and historical market return from the aggregated data in 19_expectations_analysis.do to estimate vector autoregression of moments of distribution of beliefs and plot the impulse response.

Produce the following exhibits:

- Figure 7

helper.cpp: Some functions written in rcpp to speed up computation. This is for the simulated ML method in 10_etf_demand_est_sml.R.

References

- AII (2019). American Association of Individual Investors (AII) Investor Sentiment Survey. <https://www.aaii.com/sentimentsurvey>.
- Baker, S. R., N. Bloom, S. J. Davis, and K. J. Kost (2019). Policy news and stock market volatility. *NBER Working Paper*.
- Bloomberg L.P. (2020). ETFs and volatility indices from 2008 to 2020, Retrieved from Bloomberg Terminal.
- Board of Governors of the Federal Reserve System (2020). Retail money market funds [rmfsl], retrieved from FRED, Federal Reserve Bank of St. Louis. <https://fred.stlouisfed.org/series/RMFSL>.
- CRSP Mutual Fund (2020). Center for Research in Security Prices (CRSP), The University of Chicago Booth School of Business, retrieved from Wharton Research Data Services.
- Dong, E., H. Du, and L. Gardner (2020). An interactive web-based dashboard to track covid-19 in real time. *The Lancet infectious diseases* 20(5), 533–534.
- ETF Global (2020). Retrieved from Wharton Research Data Services.
- Graham, J. R. and C. R. Harvey (2011). The Duke/CFO business outlook survey. *Duke CFO Global Business Outlook*. <http://www.cfosurvey.org>.
- Kenneth R. French - Data Library (2019). https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html.
- Lettau, M. and S. Ludvigson (2001). Consumption, aggregate wealth, and expected stock returns. *The Journal of Finance* 56(3), 815–849.
- Ludvigson, S., S. Ma, and S. Ng (2020). Uncertainty and business cycles: Exogenous impulse or endogenous response? *American Economic Journal: Macroeconomics* forthcoming.
- Martin, I. (2017). What is the expected return on the market? *The Quarterly Journal of Economics* 132(1), 367–433.
- OptionMetrics (2020). Retrieved from Wharton Research Data Services.
- Shiller, R. J. (1992). *Market volatility*. MIT press.
- Survey of Consumers (2019). University of Michigan. <https://data.sca.isr.umich.edu/>.
- Survey of Professional Forecasters (2019). Federal Reserve Bank of Philadelphia. <https://www.philadelphiafed.org/surveys-and-data/real-time-data-research/survey-of-professional-forecasters>.
- United States Stock Market Confidence Indices (2019). Yale School of Management, International Center for Finance. <https://som.yale.edu/faculty-research-centers/centers-initiatives/international-center-for-finance>.
- Wells Fargo/Gallup (2019). Investor and Retirement Optimism Index. https://news.gallup.com/topic/investor_and_retirement_optimism_index.aspx.
- Yahoo Finance (2020). S&P 500 and VIX indices. <https://finance.yahoo.com/>.