MICHIGAN ROSS

CHARACTERISTICS REGRESSION

Jason Hall, B.Com (Hons) CFA PhD
uqjhall@umich.edu
USA +1 734 926 6989
Australia +61 419 120 348

January 28, 2023

Contents

Tables

Introduction

- The objective of those note is to illustrate the use of ordinary least squares (OLS) regression to identify stock characteristics associated with above-benchmark returns.
- This note is intended for teaching purposes. Your report should be written in complete paragraphs, not bullet points.

Method and scope

- Regression of individual stock abnormal returns against stock characteristics and indicator (1, 0) variables by industry.
- Monthly returns on companies that at any stage formed part of the S&P 400 Midcap index.
- Benchmark returns on the S&P 400 Midcap index.
- Industry definitions derived from FactSet Economic Sectors and FactSet Industry Sectors.
- Monthly returns constrained at the upper and lower ends of the distribution such that monthly returns are capped at 100 per cent (adjusted monthly returns). The returns distribution is constrained at the lower 0.07th percentile (-59 per cent) and the upper 99.93rd percentile (100 per cent) across all observations. I refer to the constrained returns as adjusted stock returns. The constraint is important because otherwise a small number of observations can be unreasonably influential (for example, Gamestop recorded a return of 1,625 per cent in January 2021).
- Comparison of results based upon unadjusted and adjusted monthly returns, and with inclusion and exclusion of companies associated with high Cook's D influence statistics.
- Discussion of statistical significance and the distribution of error terms.

Data

- **Observations.** 1,471 companies and 309,269 company-month combinations that met the following criteria.
  - At any stage from December 1994 were included in the S&P 400 Midcap index.
  - Monthly data available on returns, prior month's share price, prior month's number of shares on issue, FactSet Economic Sector and FactSet Industry Sector.
  - At least 24 months of returns available for months ending January 1995 to March 2022.
- **Descriptive statistics.** Table 1 and Table 2 summarize the sample in terms of industry breakdown and the distribution of returns and market capitalization.
  - There are 24 FIN 427 Industries, with 21 of these industries having at least 2 per cent of sample observations. The most represented industry is 1300 Electronic Technology with 10.0 per cent of observations and 9.8 per cent of companies.

Table 1. Distribution of companies and observations across industries

		Number		Percentage	
Group	FIN 427 Industry	Observations	Companies	Observations	Companies
6	1100 Non-Energy Minerals	6,792	30	2.2	2.0
100	1200 Producer Manufacturing	18,366	79	5.9	5.4
7	1300 Electronic Technology	30,790	144	10.0	9.8
11	1400 Consumer Durables	10,305	42	3.3	2.9
2	2100 Energy Minerals	11,371	57	3.7	3.9
14	2200 Process Industries	12,831	62	4.1	4.2
15	2300 Health Technology	23,382	110	7.6	7.5
100	2400 Consumer Non-Durables	13,664	65	4.4	4.4
100	3100 Industrial Services	11,456	51	3.7	3.5
4	3200 Commercial Services	12,551	62	4.1	4.2
1	3250 Distribution Services	7,044	35	2.3	2.4
9	3300 Technology Services	24,185	129	7.8	8.8
16	3350 Health Services	7,858	43	2.5	2.9
8	3400 Consumer Services	17,732	88	5.7	6.0
10	3500 Retail Trade	17,610	91	5.7	6.2
12	4600 Transportation	7,923	37	2.6	2.5
100	4700 Utilities	13,706	66	4.4	4.5
5	4801 Banks	18,460	81	6.0	5.5
100	4802 Finance NEC	12,219	59	4.0	4.0
3	4803 Insurance	10,720	48	3.5	3.3
100	4885 Real Estate Dev	407	2	0.1	0.1
13	4890 REIT	17,525	73	5.7	5.0
100	4900 Communications	2,168	16	0.7	1.1
100	6000 Miscellaneous	204	1	0.1	0.1
		309,269	1,471	100.0	100.0

- Typical of samples of stock returns, the adjusted stock returns exhibit skewness. The average adjusted stock return is 1.5 per cent compared to the median of 1.2 per cent. 75 per cent of monthly stock returns lie within the range of -9.9 per cent to 12.6 per cent. The average monthly return on the S&P 400 Midcap index is -1.1 per cent and 75 per cent of benchmark returns lie within the range of -4.3 per cent to 6.0 per cent. We have preliminary evidence that small stocks earn higher returns than large stocks, given the mean adjusted abnormal return of 0.4 per cent. The mean returns in the table are equal-weighted, and the S&P 400 Midcap index is a value-weighted index. So, allocating funds based on market capitalization generated an average monthly return which was 0.4 per cent lower than allocating funds on an equal-weighted basis.
  - Market capitalization is transformed to its natural logarithm of thousands of dollars (for example, if market capitalization is $2 billion, LN 2,000,000 = 14.5). If there is a relationship between company size and returns, it is more likely to be apparent in the log transformation. Market capitalization in the S&P Midcap 400 is skewed towards a small number of large firms, typical of any index.

- The five largest stocks in the sample, on average, are QUALCOMM (1300 Electronic Technology, $68 billion). NVIDIA (1300 Electronic Technology $58 billion), WorldCom (4900 Communications $55 billion), Dell (1300 Electronic Technology $51 billion) and Netflix (3300 Technology Services $50 billion).
- The five smallest stocks in the sample have average market capitalization of less than $135 million.

Table 2. Distribution of stock returns, abnormal returns and benchmark returns

	Mean	12.5	Median	87.5
Adjusted stock return	1.5%	-9.9%	1.2%	12.6%
S&P 400 Midcap	1.1%	-4.3%	1.5%	6.0%
Adjusted abnormal return	0.4%	-9.7%	0.0%	10.3%
Market capitalization ($b)	4.4	0.4	2.0	7.9
LN market capitalization	14.5	13.0	14.5	15.9

Results

Small companies earn higher returns than large companies

- The dependent variable is abnormal returns on a given stock in a given month, with abnormal returns computed as stock returns minus benchmark returns.
- Our objective is to determine which persistent stock characteristics are associated with above-benchmark returns so we can form a portfolio weighted heavily towards those characteristics.
- So, we will test whether a stock's natural logarithm of market capitalization at the start of the month is associated with comparatively higher stock returns. This means the first independent variable is the natural logarithm of market capitalization of the stock at the start of the month.
- Stock returns are primarily determined by news associated with the economy, a company's industry and the company itself. Stock returns will therefore not be highly predictable according to a persistent characteristic like market capitalization. This means that the R-squared from the regression will be low. How implausible would it be to be able to predict stock returns with a high degree of certainty merely by computing share price times number of shares, given that all investors have access to this free information? However, it might be the case that, on average, stock characteristics are associated with above-benchmark returns. So, if we can diversify our portfolio across industries and companies, we can increase our exposure to the returns signal (for example, size) and mitigate exposure to industry- and company-specific risks. This would allow us to, on average, earn returns that exceed benchmark returns without exposing the portfolio to undue risk.
- We don't want to pick up a spurious relationship between size and returns, if indeed some industries happened to perform well over our sample period, and either large or small firms are concentrated in those industries. So, we include industry indicator (1, 0) variables to control for average industry abnormal returns over our sample period.
- Expressed as an equation our regression model is given below. There is no intercept because there is a coefficient for each industry. We could instead have included an

intercept and left out one industry indicator variable, and the interpretation of each industry coefficient would have been the incremental industry return, relative to whichever industry was left out of the indicator variables.

Return on company i in month t minus S&P 400 Midcap return in month t
= Coefficient 1 × LN Market capitalization
+ the sum of 24 coefficients on industry indicator variables × 24 industry indicator variables
+ an error term for company i in month t

- Regression results are presented in Table 3.
  - Industry indicator variables alone can explain 0.19 per cent of the variation of abnormal stock returns across time and across companies.
  - Small companies, on average, earn higher returns than large companies. The natural logarithm of market capitalization is inversely related to abnormal returns with a coefficient of -0.00479. A one unit difference in a company's natural logarithm of market capitalization is associated with a 0.48 per cent difference in monthly abnormal returns in the opposite direction. To put this in perspective, for the natural logarithm of market capitalization the 25th percentile is 13.7 and the 75th percentile is 15.3. The size reduction from the 75th percentile to the 25th percentile is associated with an approximate 0.8 per cent increase in average monthly returns, equivalent to about 9.2 per cent a year.

Table 3. Regression results

Model	Dependent variable = Adjusted abnormal returns	Dependent variable = Adjusted abnormal returns	Dependent variable = Unadjusted abnormal returns
LN MC Coefficient		-0.00479	-0.00529
LN MC Std error (assuming errors are independent and error variance is constant)		0.00016	0.00017
LN MC Std error (adjusted for non-constant error variance)		0.00021	0.00023
LN MC Std error (adjusted for repeated observations by company)		0.00023	0.00025
Industry indicators	Yes	Yes	Yes
R-squared (%)	0.06	0.35	0.37
Adj r-squared (%)	0.05	0.34	0.36
DW	2.04	2.03	2.04
N = 309,269			

Constrained variables and influential observations

- If we ran the regression without constraining monthly returns within the range of -59 per cent to +100 per cent, the coefficient on the natural logarithm of market capitalization would have been -0.00529, an increase of 10 per cent. This is material because it corresponds an annual increase in predicted returns of 1.0 per cent for the difference between a stock at the 25th percentile and 75th percentile of the size distribution.

- □ For stock returns, it was easy to see that a handful of returns were very large, and we could easily make that case that we cannot reasonably form a portfolio on the basis of results from a small number of outsized returns.
- □ The Cook's D influence statistic is a measure of the relative influence of an observation on the regression output. When I ran the regression using abnormal returns without constraints, the observation with the highest Cook's D was Gamestop in January 2021 with 1625 per cent return. If we run the regression again, using unadjusted abnormal returns, and remove this single observation out of over 300,000 the coefficient on LN MC falls to -0.00527. If we run the regression and remove the 100 most influential observations, the coefficient on LN MC falls to -0.00460. The corresponding regression coefficient using adjusted returns with the same small number of observations removed is -0.00448.
- □ **In short, if you retain extreme returns in your data without adjustment, those returns will be highly influential to your conclusions and you are likely to bias your portfolio formation decisions towards stock characteristics that, by chance, were associated with a small number of extreme returns in an historical series that is unlikely to be repeated.**
- □ Sometimes, the Cook's D influence statistic suggests that an entire company is highly influential to the results. For example, Wisdomtree Investments Inc (CUSIP = 97717P10), a provider of exchange-traded funds (ETFs) appears 20 times in the top 50 observations based upon Cook's D influence statistics using the unadjusted abnormal returns. If this entire company is removed, along with a single month from Gamestop, the coefficient on LN MC using unadjusted abnormal returns falls to -0.00527 and the coefficient on LN MC using adjusted abnormal returns fall to -0.00479.
- □ **The point is that in portfolio management, the magnitude of relationships between stock characteristics and returns matters, not just direction and statistical significance. We always need to check for possible spurious relationships in the data, regardless of whether or not there is what appears to be a statistically significant relationship between variables in a dataset.**

Statistical significance

- □ Examining the coefficient on LN Market Capitalization and its standard error, the OLS regression results strongly suggest statistical significance. Statistical significance is the probability of observing a result different to our null hypothesis due to the sample being non-representative of the population. The population is all potential stock returns on all potential mid-sized companies that could have been listed. Imagine history was repeated over and over again in parallel universes in which different companies emerged, and they exhibited different stock returns when listed. We only get to observe one sample of 1,471 companies for one 28-year time period from January 1995 to March 2022. We can create a distribution of alternative samples by repeatedly drawing from our original sample, as an approximation of the samples that we could have drawn from the unobservable population. That does not form part of our course but is a more accurate way of estimating statistical significance than drawing inferences from the standard error in regression.

- Using adjusted abnormal returns, the OLS standard error is 0.00016. The t-statistic for testing whether the coefficient of -0.00467 is different from a null hypothesis of zero is (-0.00467 – 0)/0.00016 = -29.5, which implies a p-value of less than 0.01 per cent. However, you should be aware that the standard error from OLS regression is understated in this instance. This will not form a material part of our course, but is something you should be aware of.
    - The standard error generated by OLS regression is based upon the assumption that the errors are **independent** and have the **same variance** regardless of characteristics of the independent variables. The error is the difference between the actual adjusted abnormal return, and the predicted adjusted abnormal return from the regression. In OLS regression, there is an assumption that whether the error is positive or negative and whether it is a far from or close to the predicted value, has nothing to do with the companies in the dataset, their market capitalization, and any other error terms. The technical term for error terms having the same variance regardless of the independent variables is homoskedasticity (meaning constant variance). The technical term for non-constant variance is heteroskedasticity.
    - The error terms do not exhibit constant variance across observations. One correction that I applied to compute standard errors that adjust for non-constant variance generates a standard error of 0.00021. This means there is a roughly 25 per cent increase in the standard error once non-constant variance in error terms is corrected.
    - The error terms are also not independent across observations because the same company appears multiple times in the dataset. If a company made consistent positive announcements over the sample period, it will exhibit consistent returns above what is predicted by the regression model, so the error terms will, in part, be dependent upon the individual company. I applied a correction called clustered standard errors, clustering by company, to account for this. The standard error is now 0.00024, a 50 per cent increase above the baseline standard error of 0.00016.
    - In summary, tests of statistical significance are always based upon a suite of underlying assumptions, and violation of these assumptions can potentially lead to spurious conclusions about statistical significance. In our case, this is not a problem because even with clustered standard errors, the sample size of over 300,000 months, almost 1,500 companies and 27 years of data, makes it unlikely that small firm returns exceeded large firm returns purely by chance. This is shown by the comparatively small standard error, even when computed using clustered standard errors.

Stock characteristics worth of consideration

- We know that company size is a stock characteristic worthy of consideration in portfolio formation. Researchers from the 1980s onwards documented that, on average, small stocks earn higher returns than large stocks (for example, Fama and French, 1992).
- Now we are in a position where you should suggest stock characteristics that are plausibly related to stock returns. Another obvious candidate is the book-to-market equity ratio (book value of equity scaled by market value of equity), also used by Fama and French (1992).

- □ The Invesco S&P MidCap Quality ETF is based on the S&P MidCap 400 Quality Index. The list below outlines the construction of the quality score used to determine the 80 stocks that enter into the S&P Midcap 400 Quality Index.[1]
  - "Constituent selection. The top 80 eligible securities are selected for the index by quality score. This score is based on a composite of the following three factors [which we call characteristics] for each stock. Individual z-scores are calculated and then combined into one single metric, which is then used to rank constituents. Each of the individual factor values are winsorized. Z-score is (observation value – average)/standard deviation, in other words, the number of standard deviations from the mean.
    - Return on equity: ... Trailing 12-month earnings per share divided by the latest book value per share. If either earnings per share or book value of equity is negative, the observation is assigned the lowest return on equity z-score. Observations are constrained at the 2.5th and 97.5th percentiles.
    - Accruals ratio [this is mislabeled as "accruals" by S&P as the metric captures asset growth and not accounting accruals"]: ... Change in net operating assets over the last year divided by average net operating assets over the last two years. Observations are constrained at the 2.5th and 97.5th percentiles.
    - Financial leverage: ... Latest total debt divided by book value of equity. If book value of equity is negative the observation is assigned the lowest financial leverage z-score. Observations are constrained at the 2.5th and 97.5th percentiles/
  - Constituent weighting. Components are weighted by a product of the quality score assigned and the float-adjusted market capitalization.
- □ In summary, there are five stock characteristics that we know could form part of our quantitative investment strategy. **You will now attempt to come up with other characteristics and your rationale for those characteristics.**
  - Size: Smaller implies higher returns
  - Book-to-market ratio: Larger implies higher returns
  - Return on equity: Larger implies higher returns
  - Accruals ratio: Smaller implies higher returns
  - Financial leverage: Smaller implies higher returns
- □ These characteristics may be positively correlated. For example, companies with a higher book-to-market ratio could exhibit lower return on equity (more investment needed to generate profits, lower return on investment), and companies with a higher book-to-market ratio may have higher leverage (utilities have higher leverage than technology companies, and a higher proportion of market value represented by equity on the balance sheet). As we expand the candidates for drivers of portfolio composition, we need a mechanism to retain the most important characteristics and discard others. So, next we will analyze a suite of characteristics using both OLS and LASSO regression, with the latter including a penalty term for additional explanatory variables, and a response that sets coefficients equal to zero for variables that are considered to redundant, given other variables available for analysis.

[1] Factsheet and detailed methodology available at https://www.spglobal.com/spdji/en/indices/strategy/sp-midcap-400-quality-index/#overview, accessed on January 29, 2023/

References

Fama, E.F., and K.R. French, 1992. The cross-section of expected stock returns, Journal of Finance, 47, 427-465.