

Problem Set: Linear Regression

Michael R. Roberts

August 16, 2010

I. Getting to Know Compustat

Replicate the summary statistic tables attached to the back of this problem set. The sample consists of firm observations with 1990 fiscal year end, and excludes all financial firms, utilities, foreign governments, international affairs, non-operating establishments, and companies not headquartered in the United States. See the attachment below for help with these screens and variable definitions.

II. Cross-Sectional Univariate Linear Regression

Using the sample described above, estimate the following cross-sectional regression,

$$Investment_i = \beta_0 + \beta_1 MB_i + u_i,$$

where $Investment_i$ is the ratio of capital expenditures in 1990 to net PPE in 1989, MB is the market to book ratio, and i indexes firms. Before estimating, winsorize all variables at the upper and lower one percentiles. See the attachment below for more information on variable definitions.

1. Interpret each number in the regression output. Specifically, show how each number is calculated and what it tells us.
2. What is the mean of the estimated residuals? Is this surprising?
3. What is the correlation between the estimated residuals and the market-to-book ratio? Is this surprising?
4. What are the means of the predicted values and the actual investment values?
5. Produce a scatter plot of the data with an overlay of the fitted regression line. Are we missing any nonlinearity? Does the picture suggest heteroskedasticity?
6. Re-estimate the model and compute heteroskedasticity consistent standard errors. How do they compare to the OLS estimators?
7. Re-estimate the model using the lagged value of the market-to-book ratio. How do these estimate compare to that of the contemporaneous value?
8. Estimate the constant elasticity version of the model. Interpret the coefficients. Does the sample size change? If so, why? How does the R^2 compare with the levels version?
9. Estimate the Log-level version of the model. Interpret the coefficients. Does the sample size change? If so, why? How does the R^2 compare with the levels and constant elasticity versions?
10. What does slope the coefficient tell us about the q-theory? Do we have a clean test of q-theory with this data and specification?

III. Cross-Sectional Multivariate Linear Regression

A. Investment Regression

Using the sample described above, estimate the following cross-sectional regression,

$$Investment_i = \beta_0 + \beta_1 MB_i + \beta_2 CashFlow_i + u_i$$

where $Investment_i$ is the ratio of capital expenditures in 1990 to net PPE in 1989, MB is the market to book ratio, $CashFlow$ is the ratio of cash flow in 1990 to net ppe in 1989, and i indexes firms. Before estimating, winsorize all variables at the upper and lower one percentiles. See the attachment below for more information on variable definitions.

1. Estimate the model and interpret your results (coefficients, t-stats, model fit).
2. Restimate the model only this time use lagged values for the independent variables, market-to-book and cash flow. (i.e., market-to-book in 1989 and cash flow in 1989 normalized by net ppe in 1988). Is there any change from 1.? If so, why?
3. Replace $CashFlow$ with a quadratic market-to-book term. Discuss and interpret your findings. Construct a figure containing the fitted curve and a scatter plot of the data. Discuss.
4. Test for heteroskedasticity in the level-level specification using lagged independent variables?
5. Rerun the regression with lagged independent variables using heteroskedasticity robust standard errors. Does this make any difference for inference?
6. Using the lagged independent variable specification, test the null hypothesis that the two slope coefficients are equal. Be sure to use the heteroskedasticity consistent standard errors. Perform this test two ways. First, conduct a standard F-test. Second, reparameterize the model so that the regression outputs a simple t-statistics corresponding to the null.
7. Test the null hypothesis that the sum of the two slope coefficients equals 0.03. Perform this test two ways. First, conduct a standard F-test. Second, reparameterize the model so that the regression outputs a simple t-statistics corresponding to the null.

8. Reestimate using a log-level specification (lagged independent variables)? How does the fit compare with the level-level? What happens to the # of observations?
9. Reestimate using a log-log specification (lagged independent variables)? How does the fit compare with the log-level and level-level specifications? What happens to the # of observations?
10. Discuss the economic interpretations of the results, as well as concerns you might have with the interpretations.

B. Leverage Regression

Using the sample described above, estimate the following cross-sectional regression,

$$Leverage_i = \beta_0 + \beta_1 Profitability_i + \beta_2 Tangibility_i + \beta_3 FirmSize_i + \beta_4 MB_i + u_i$$

where *Leverage* is the ratio of total debt to total assets, *Profitability* is the ratio of operating income before depreciation to total assets, *Tangibility* is the ratio of net PPE to total assets, *FirmSize* is the natural logarithm of total assets, *MB* is the market to book ratio, and *i* indexes firms. Before estimating, winsorize all variables at the upper and lower one percentiles. See the attachment below for more information on variable definitions.

1. Estimate the model with contemporaneous independent variables and interpret your results (coefficients, t-stats, model fit).
2. Reestimate the model only this time use lagged values for the independent variables, market-to-book and cash flow. (i.e., market-to-book in 1989, etc.) Is there any change from 1.? If so, why?
3. Are OLS standard errors appropriate? Or, should we be concerned with heteroskedasticity?
4. Reestimate the model using heteroskedasticity consistent SEs. Did anything change?
5. Which regressor has the largest partial effect on leverage? I.e., which regressor appears to be the most economically important determinant?
6. Test the null hypothesis that all of the slope coefficients are equal.

7. Test the null hypothesis that 2 times the slope on $\log(\text{assets})$ less 0.5 times the slope on tangibility equals -0.06.
8. Discuss the economic interpretations of the results, as well any concerns you might have with the interpretations.

IV. Cross-Sectional Multivariate Linear Regression - Dummy Variables

We're going to introduce credit ratings into this exercise. To get the ratings, you'll need the dataset "adsprate," which can be found at "/wrds/comp/sasdata/na/rating." You'll need to construct a calendar year variable from the "datadate" variable and then crunch the data down to unique gvkey-year observations by picking off the last observation within a gvkey-year combination. Then merge this data with the compustat data by gvkey-year combination. Be careful to merge using calendar year from compustat to get the alignment correct. For that, you'll need the datadate variable from compustat to determine the calendar year of the filing.

The relevant rating variable is "splticrm." A firm is deemed to be rated if it's rating is a nonmissing, letter rating. Ratings equal to "N.M.", "SD", and "Suspended" are not valid ratings. Investment grade debt is rated "BBB-" or higher, speculative grade debt is rated lower than "BBB-."

Using the same 1990 dataset from above with the newly appended credit rating information answer the following questions. Before estimating, winsorize all continuous ratio variables at the upper and lower one percentiles. *Leverage* is the ratio of total debt to total assets in 1990, *CreditRatin* is an indicator variable equal to one if the firm has a long-term issuer credit rating from S&P as of 1989.

1. Regress book leverage on a credit rating indicator (1 if there is a valid credit rating, 0 otherwise) using OLS and heteroskedasticity robust standard errors. Interpret the results statistically and economically.
2. Incorporate firm size, profitability, tangibility, and market-to-book and re-estimate the model. Discuss any changes in the credit rating coefficient.
3. Center all of the continuous variables at their sample means (i.e., subtract the mean value from each observation). Rerun the regression in 2 and discuss the results. What changed/did not change? Why?
4. Create indicator variables for each major rating level (i.e., AAA, AA, A, BBB, BB,...) and tabulate the the number of observations per each rating level.
5. Regress leverage on these rating level indicators using only the subsample of rated firms. Interpret the results. Perform a joint test of whether the coefficients across

all ratings levels are statistically different. Perform a joint test of whether leverage varies significantly within each letter rating (e.g., AAA = AA = A).

6. Create an indicator variable identifying high and low growth firms based on above and below (or equal to) median market-to-book ratios, respectively. Regress leverage on the high growth indicator. Then regress leverage on the high growth indicator, the credit rating indicator from 1. and the interaction.
 - Interpret each coefficient.
 - What is the average leverage ratio for each of the four combinations of rated and growth?
 - Confirm your interpretations and analysis by computing simple averages for the relevant subsamples defined by the conditional expectations.
7. Regress leverage on the credit rating indicator, firm size, profitability, tangibility, market-to-book, and market-to-book interacted with the credit rating indicator. Then rerun the same regression after centering each continuous variable around its sample mean. Compare the results and discuss.

V. Data

The data is from the annual Compustat database, FUNDA, and is located on WRDS at `"/wrds/comp/sasdata/na"`. In the investment and capital structure literatures, a variety of screens are frequently used to eliminate certain observations from the sample. A few of these screens include the following.

1. `(indfmt == "INDL" & datafmt == "STD" & popsrc == "D" & consol == "C")`: These conditions ensure that `gvkey-datadate` uniquely identify each observation. (`gvkey-fyear` is "almost" the unique identifier, but for 48 obs.)
2. `(year ≥ 1965)`: Observations with year-ends greater than or equal to 1965. Prior to this year, selection issues become particularly severe in Compustat.
3. `(sic ≥ 0000 & sic ≤ 999)`: Agriculture, Fishing & Hunting.
4. `(sic ≥ 4900 & sic ≤ 4999)`: Utilities.
5. `(sic ≥ 6000 & sic ≤ 6999)`: Financial Firms.
6. `(sic = 8888)`: Foreign Governments.
7. `(sic ≥ 9000 & sic ≤ 9999)`: International affairs & non-operating establishments.
8. `(gvkey != "")`: No Missing company indicators.
9. `(fyear != .)`: No Missing fiscal year indicators.
10. `(fic == "USA")`: Only firms headquartered in the United States.
11. `(prcc.f = . & csho = .)`: Nonmissing stock market data (price & shares outstanding).

The investment variable definitions come from Hennessy, Levy, & Whited 2007 (JFE). The capital structure variable definitions come from Lemmon, Roberts, and Zender (2008) (JF). (In the construction of the market-to-book ratio here, you must first set all missing observations for `pstkl` and `txditc` to zero.)

Investment Variables			
g	mb	= (at + (prcc.f * csho) - ceq - txdb) / at;	(Market-to-Book)
g	cf	= (ib + dp);	(Cash flow)
g	cf_k	= cf / ppent[_n-1];	(Cash flow / capital(t-1))
g	inv	= (capxv - sppe);	(Net Investment)
g	inv_k	= inv / ppent[_n-1];	(Investment / capital(t-1))

Capital Structure Variables			
g	td	= dlc + dltd;	(Total Debt)
g	bl	= td / at;	(Book leverage)
g	ml	= td / (td + (prcc.f * csho));	(Market leverage)
g	prof_a	= oibdp / at;	(Profitability)
g	tang_a	= ppent / at;	(Tangibility)
g	me	= prcc.f * csho;	(Market equity)
g	mk2bk	= (me + dlc + dltd + pstkl + txditc) / at;	(Market-to-book)
g	loga	= log(at);	(Log(assets))
g	logsale	= log(sale);	(Log(sales))
g	zscore	= (3.3 * pi + sale + 1.4 * re + 1.2 * (act - lct)) / at;	(Altman's unlevered Z-score)

Table I
Summary Statistics

The sample consists of all U.S., nonfinancial, nonutility firms in the annual Compustat database in fiscal year 1990. Also excluded are foreign governments, international affairs, non-operating establishments, and companies not headquartered in the United States. Panel A presents summary statistics for the raw data. Panel B presents summary statistics for the data after trimming each variable at the upper and lower one percentiles. Panel C presents summary statistics for the data after winsorizing each variable at the upper and lower one percentiles.

Panel A: Raw Data

	N	Mean	SD	Min	1	5	25	50	75	95	99	Max
Investment / Capital	4,538	10.55	605.66	-33.00	-0.39	0.00	0.08	0.18	0.38	1.38	7.75	40737.00
Market-to-Book (Inv)	4,482	1.85	5.39	0.17	0.46	0.66	0.93	1.17	1.74	4.49	10.71	315.07
Cash Flow / Capital	5,641	2.96	240.27	-847.00	-25.20	-4.30	-0.06	0.17	0.45	2.16	11.29	17846.00
Book Leverage	6,020	0.39	2.18	0.00	0.00	0.00	0.09	0.28	0.45	0.86	1.82	151.86
Market Leverage	4,702	0.34	0.29	0.00	0.00	0.00	0.07	0.29	0.55	0.87	0.98	1.00
Profitability	6,006	-0.05	2.97	-159.50	-1.84	-0.50	0.01	0.10	0.17	0.28	0.43	14.50
Tangibility	6,042	0.34	0.26	0.00	0.00	0.02	0.12	0.27	0.51	0.85	0.93	1.00
Market-to-Book (CS)	4,702	1.56	5.15	0.00	0.10	0.29	0.68	0.95	1.48	4.13	10.09	306.79
Log(Assets)	6,121	4.11	2.55	-6.91	-1.19	0.24	2.19	4.02	5.92	8.39	9.61	12.10
Z-Score	5,528	-3.01	118.39	-6933.90	-32.98	-7.46	0.25	1.49	2.58	4.00	5.71	72.35

Panel B: Trimmed Data

	N	Mean	SD	Min	1	5	Percentiles							Max
							25	50	75	95	99			
Investment / Capital	4,448	0.36	0.68	-0.39	-0.11	0.00	0.08	0.18	0.37	1.22	3.73	7.75		
Market-to-Book (Inv)	4,394	1.62	1.31	0.46	0.54	0.69	0.93	1.17	1.73	4.09	7.76	10.71		
Cash Flow / Capital	5,529	-0.17	2.60	-25.20	-12.79	-3.51	-0.04	0.17	0.44	1.84	4.93	11.29		
Book Leverage	5,960	0.31	0.27	0.00	0.00	0.00	0.09	0.27	0.45	0.81	1.20	1.82		
Market Leverage	4,655	0.33	0.28	0.00	0.00	0.00	0.07	0.29	0.54	0.86	0.94	0.98		
Profitability	5,886	0.04	0.25	-1.84	-1.15	-0.41	0.01	0.10	0.16	0.27	0.37	0.43		
Tangibility	5,922	0.34	0.25	0.00	0.01	0.02	0.13	0.27	0.51	0.83	0.90	0.93		
Market-to-Book (CS)	4,608	1.34	1.26	0.10	0.14	0.34	0.68	0.95	1.46	3.76	7.05	10.09		
Log(Assets)	5,999	4.11	2.41	-1.19	-0.61	0.38	2.23	4.02	5.87	8.23	9.23	9.61		
Z-Score	5,418	0.62	3.88	-32.98	-17.63	-6.18	0.30	1.49	2.55	3.88	4.92	5.71		

Panel C: Winsorized Data

	N	Mean	SD	Min	1	5	Percentiles							Max
							25	50	75	95	99			
Investment / Capital	4,538	0.42	1.00	-0.39	-0.39	0.00	0.08	0.18	0.38	1.38	7.75	7.75		
Market-to-Book (Inv)	4,482	1.70	1.59	0.46	0.46	0.66	0.93	1.17	1.74	4.49	10.71	10.71		
Cash Flow / Capital	5,641	-0.30	3.76	-25.20	-25.20	-4.30	-0.06	0.17	0.45	2.16	11.29	11.29		
Book Leverage	6,020	0.33	0.31	0.00	0.00	0.00	0.09	0.28	0.45	0.86	1.82	1.82		
Market Leverage	4,702	0.34	0.29	0.00	0.00	0.00	0.07	0.29	0.55	0.87	0.98	0.98		
Profitability	6,006	0.03	0.31	-1.84	-1.84	-0.50	0.01	0.10	0.17	0.28	0.43	0.43		
Tangibility	6,042	0.34	0.26	0.00	0.00	0.02	0.12	0.27	0.51	0.85	0.93	0.93		
Market-to-Book (CS)	4,702	1.42	1.53	0.10	0.10	0.29	0.68	0.95	1.48	4.13	10.09	10.09		
Log(Assets)	6,121	4.12	2.50	-1.19	-1.19	0.24	2.19	4.02	5.92	8.39	9.61	9.61		
Z-Score	5,528	0.34	5.11	-32.98	-32.98	-7.46	0.25	1.49	2.58	4.00	5.71	5.71		