Monte Carlo Simulation Analysis of Fama–French Five-Factor Model

Team Member:

HU WEI **54206763**

HUANG Xiaoming **54934428**

WANG Qiping **54486639**

Abstract

Linear Regression with O.L.S method is most popular approach for modelling the relationship between variables in academic and practical fields. However, researchers and practitioners often misuse the O.L.S properties by ignoring the prerequisites. In this paper, we use Monte Carlo simulations to validate the Fama–French five-factor model and the O.L.S properties. We confirm the validity of five factors in the Fama–French model, and uncover six interesting theoretical findings of O.L.S properties.

1 Introduction

The Fama-French model is a model widely adopted in portfolio management and asset pricing. It was initially designed by Eugene Fama and Kenneth French in 1993 to capture stock returns. Overcoming the drawbacks of the Capital Asset Pricing Model (CAPM) that merely uses one market risk factor to describe portfolio or stock returns, the Fama–French three-factor model extends it by adding two additional risk factors related to size and Book-to-Market ratio[1]. These factors can reflect specific features about firms. Fama and French (1992) demonstrated that size and book-to-market ratio are connected with economic fundamentals, and can act as proxies for risk factors in stock returns. The model well reflects the fact that small-cap and value stocks outperform the whole market. The effectiveness of the model has been widely demonstrated by both industry and academic fields [2, 3]. In 2014, Fama and French expanded the original three-factor model by adding two additional factors namely profitability and investment patterns, thus generating a five-factor model [4]. Compared with the three-factor model, the five-factor model takes the variation in average returns regarding to profitability and investment into consideration. In this study, we adopted the Fama–French five-factor model as our target model.

Monte Carlo methods, often applied in mathematical and physical domains, are a set of computational algorithms that acquire numerical results through random sampling. Monte Carlo methods are often targeted for addressing three classes of problems, including simulation and optimization, probability distribution and numerical integration [5].

In this project, we mainly focus on the utilization of Monte Carlo simulations to test the validity of the Fama–French five-factor model. More specifically, we downloaded the factors (R_m-R_f), SMB (Small [market capitalization] Minus Big), HML (High [book-to-market ratio] Minus Low), RMW (Robust Minus Weak), and CMA (Conservative Minus Aggressive) ranging from Jan 1st, 2008 to Dec 29th, 2017 from the Kenneth R. French Homepage[1]. Then we got the true corresponding parameters $\beta = (\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6)^T$ according to the Fama-French five-factor model. After that, based on the assumption that the error term ε follows a standard normal distribution $\mathcal{N}(0, \sigma^2)$, we did Monte Carlo simulation experiment on ε and got the estimators of true β accordingly, which are named as $\hat{\beta} = \left(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4, \hat{\beta}_5, \hat{\beta}_6\right)^T$. Most importantly, we uncovered six interesting theoretical findings through the regression model.

The rest of the paper proceeds as follows. Section 2 involves a description of the classical Fama-French five-factor model and the empirical data set used in this project. Section 3 elaborately addresses six specific findings by utilizing the Monte Carlo simulation approach. Section 4 provides a concluding remark and highlights the directions of our future research work.

2 Model Description and Experiment Design

The Fama-French five-factor model (2014) can be expressed as Equation (1):

$$r - R_f = \beta_1 + \beta_2 \cdot \left(R_m - R_f\right) + \beta_3 \cdot \mathrm{SMB} + \beta_4 \cdot \mathrm{HML} + \beta_5 \cdot \mathrm{RMW} + \beta_6 \cdot \mathrm{CMA} + \varepsilon \quad (1)$$

And in the Equation (1),

- r is the return of target portfolio,
- R_f is the risk-free return,
- R_m is the return on the value-weighted market portfolio,
- SMB is the return on a diversified portfolio of small stocks minus the return on a diversified portfolio of big stocks,
- HML is the difference between the returns on diversified portfolio of high and low B/M stocks,

[1] http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html

- RMW is the difference between the returns on diversified portfolio of stocks with robust and weak profitability,
- CMA is the difference between the returns on diversified portfolios of stocks of low and high investment firms, and ε is a zero-mean residual.

The market excess return factor, $R_m - R_f$, captures the market information and can be proxied by value-weighted portfolio of all stocks in the market, for example, New York Stock Exchange (NYSE) in U.S. economy.

The SMB factor is constructed to measure the size premium, which is the additional return that investors have historically received by investing in stocks of firms with relatively small market capitalization. If $SMB > 0$, it means that small cap stocks have outperformed the large cap stocks at that time.

The HML factor can help us to measure the premium-value provided to investors for investing in firms with high B/M. If $HML > 0$, it suggests that "value stocks" outperform "growth stocks" at that time. Meanwhile, the book-to-market equity ratio, B/M, is the ratio of book value of the firm over the market value of the firm, where the book value of a firm is its accounting value calculated from balance sheet and the market value of a firm is determined by calculating its market capitalization, which is the product of the total number of shares outstanding multiplied by the current share price. B/M is always used to identify undervalued or overvalued securities. If $B/M < 1$, the firm is considered to be overvalued.

The RMW factor capture the profitability in average stock return and the CMA factor capture investment patterns.

2.1 Dataset Description

In the Monte Carlo simulation, we use the dataset *Fama/French 3 Factors (Daily)* and *Portfolios Formed on Book-to-Market (Daily)* download from the Kenneth R. French Homepage, which are created using CRSP database. And the risk-free return proxy is 1-month T-Bill return, which is from Ibboston and Asscociates, Inc..

The portfolios suggested by Kenneth R. French Homepage are created by following procedure: 1) Rank all stocks that trade on NYSE, Nasdaq and Amex (when it existed) by their book-to-market ratios; 2) Take the bottom 30% and put them into a portfolio (low 30%); 3) Take the next 40% and put them into a portfolio (medium 40%); 4) Take the top 30% and put them into a portfolio (top 30%); 5) Portfolio weights within each portfolio are value-weights. For our targeted portfolio for Monte Carlo Simulation, we choose the portfolio of lowest 30%, named as *lo_30*. And from the dataset, we can select the daily values for target portfolio excess return $r - R_f$, market portfolio excess return $R_m - R_f$, SMB, HML, RMW, and CMA ranging from Jan 1st, 2008 to Dec 29th, 2017 with 2,518 observations. In table 1, you can find the descriptive statistics of our dataset.

2.2 Monte Carlo Simulation Design

In Monte Carlo Simulation of linear regression, there are a few assumptions: (1) the error term ε follows a standard normal distribution $\varepsilon \sim \mathcal{N}(0, \sigma^2)$; (2) Matrix of independent variables X is non-stochastic and $cov(X, \varepsilon) = 0$; (3) X should be full column rank, which means that $rank(X) = k = 6$, where $k - 1$ is the number of independent variables. $N = 1{,}000$ is the number of replications and $M = 2{,}518$ is the number of observations.

Mooney (1997) presents five steps to make a Monte Carlo simulation. In our study of Fama-French five-factor model, we will take use of Monte Carlo method to prove some properties of OLS estimator (such as unbiasedness), confidence interval, t-test, Type I error and model selection. The Monte Carlo Simulation procedure of our study is as follow:

Step1: Run the linear regression of Fama-French five-factor model using downloaded data to get the true β;

Step 2: Choose a large replication size N and other parameters, such as significance level for t test α;

Step 3: Generate the fixed values of independent variables matrix $X \in \mathbb{R}^{M \times 6}$;

Step 4: For each replication, we'll do the same tasks as follow. In replication i, $i = 1, \dots, N$:

Step 4.1: Generate the error term ε^i under assumption (1) and set $\sigma^2 = Var(r - R_f) = 1.4060$ using downloaded dataset;

Step 4.2: Generate the values of dependent variable $(r - R_f)^i$ using Fama-French five-factor model;

Step 4.2: Make the treatment of estimating the regression parameter β by OLS method and get the OLS estimator $\hat{\beta}^i$ and estimated error variance s^{2^i};

Step 4.3: Generate the confidence interval of OLS estimator $\hat{\beta}$, CI^i, and check whether CI^i contains true β or not;

Step 4.4: Under the null hypothesis of "$\hat{\beta}^i = \beta$", check whether Type I error happens or not;

Step 4.5: Generate t-statistics under the null hypothesis "$\hat{\beta}^i = \beta$";

Step 4.6: Redo Step 4.1-4.5 for new model with omitted variables;

Step 5: Average the OLS estimator and estimated error variance, thereby forming the Monte Carlo estimates $\hat{\beta}^i$, s^{2^i};

Table 1 Descriptive Statistics and Regression Statistics

		Inter. (β_1)	Rm-Rf (β_2)	SMB (β_3)	HML (β_4)	RMW (β_5)	CMA (β_6)
Descriptive Statistics	**N**	2,518	2,518	2,518	2,518	2,518	2,518
	Mean	1	0.041	0.008	0.0002	0.0139	0.003
	St.Dev	0	1.2911	0.5957	0.7224	0.3843	0.3146
	Min	1	-8.95	-3.41	-4.22	-2.6	-1.7
	Max	1	11.35	4.48	4.83	1.94	1.97
Coefficients	**Estimate**	0.0045	0.9792	-0.0294	-0.2585	0.05524	-0.0345
	Std.Error	0.00195	0.00181	0.003511	0.003373	0.006215	0.006684
	t value	2.306	542.072	-8.374	-76.618	8.889	- 5.154
	P value	0.0212 *	< 2e-16 ***	< 2e-16 ***	< 2e-16 ***	< 2e-16 ***	2.74e-07 ***

Regression Statistics	**Multiple R**	**Adjusted R Square**	**Standard Error**		**F-statistic**		
			Value	**d.f.**	**Value**	**d.f.**	**p-value**
	0.9933	0.9933	0.09759	2512	7.486e+04	2512	< 2.2e-16

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

3 Experiment

For the step 1 of Monte Carlo procedure, we use real data as described in Section 2.1 to run the linear regression of Fama-French five-factor model and get the true β. Details about the regression and coefficients β can be found in table 1. Follow the Monte Carlo Procedure as described in Section 2.2, we will investigate some properties of OLS method from different aspect such as unbiasedness of OLS estimator, confidence interval, t-test, Type I error and model selection. And we will discuss these aspects in the following part.

3.1 Finding 1: The O.L.S. estimators of the unknown coefficients and error variance are unbiased

According to Table 1, the mean values we assign to the intercept and five independent variables ($R_m - R_f$, SMB, HML, RMW and CMA) of the Fama–French five-factor model are $\beta = (0.0045, 0.9792, -0.0294, -0.2585, 0.0552, -0.0345)^T$. Therefore, based on our dataset, we assume that the true Fama–French five-factor model can be expressed as Equation (2):

$$r - R_f = 0.0045 + 0.9792 \cdot (R_m - R_f) - 0.0294 \cdot \text{SMB} - 0.2585 \cdot \text{HML} + 0.0552 \cdot \text{RMW} - 0.0345 \cdot \text{CMA} + \varepsilon \quad (2)$$

In this study, $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4, \hat{\beta}_5, \hat{\beta}_6)^T$ is the O.L.S. estimator of $\beta = (\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6)^T$. We define $e = (e_1, e_2, e_3, e_4, e_5, e_6)^T = Y - X\hat{\beta}$. Thus $\hat{\beta} = (X'X)^{-1}X'Y$, provided that X′X′ is non-singular.

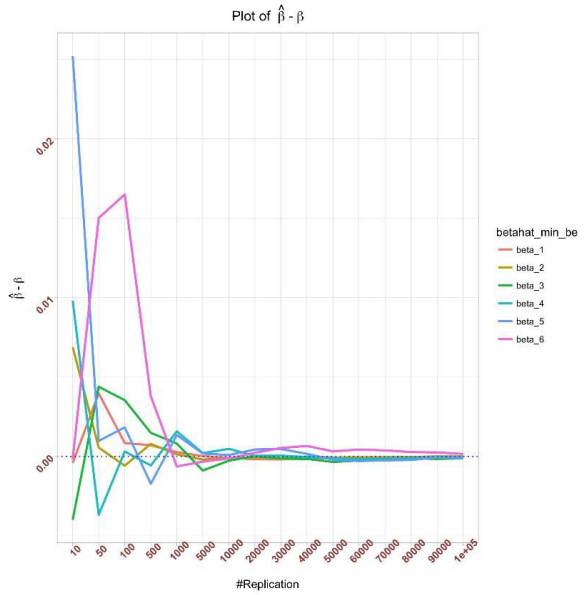

Figure 1. Convergence Graph of $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}$

Solid lines represent the values of $\hat{\beta} - \beta$ and the dashed line is the horizon line of zero value.

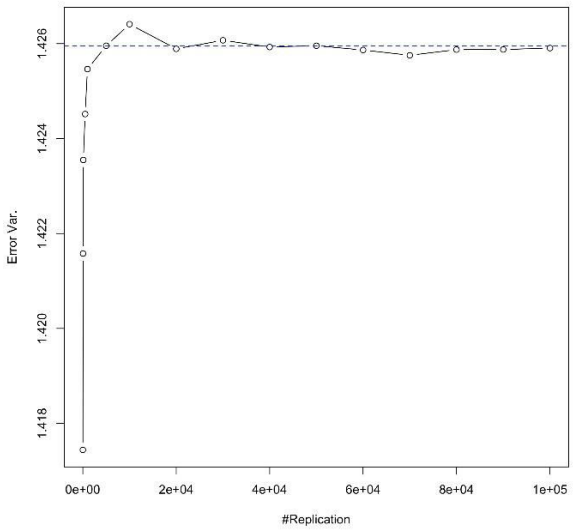

Figure 2. Convergence Graph of Error Variance

Blue dashed line represents the true error variance.

Figure 1 shows the convergence graph of $\hat{\beta} - \beta$, from which we can see that with the increase of replication numbers, all the mean values of $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4, \hat{\beta}_5$ and $\hat{\beta}_6$ converge to β_1, β_2, β_3, β_4, β_5 and β_6 respectively. It means that $\text{E}(\hat{\beta}) = \beta$ and the O.L.S estimators of β are unbiased. Besides, Figure 1 also indicates that convergence speeds vary across different estimators of coefficients. Among them, $\hat{\beta}_2$, the estimator of the market excess return factor $(R_m - R_f)$, approaches to 0 fastest and generates the best convergence effect, which means that the market excess return is the most stable factor for the market risk premium.

Error variance can be expressed as Equation (3):

$$s^2 = \frac{e'e}{n-k} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-k} \quad (3)$$

where s^2 is the O.L.S. estimator of σ^2, and $n - k$ is the model's degrees of freedom (d.o.f.).

Figure 2 shows the convergence graph of s^2, from which we can see that with the increase of replication numbers, the estimated error variance converges to the true error variance. It means that $\text{E}(s^2) = \sigma^2$ and the O.L.S. estimator of error variance is unbiased.

Therefore, we demonstrate an interesting finding that the O.L.S. estimators of the unknown coefficients and error variance of the Fama–French five-factor model are unbiased through Monte Carlo simulations.

3.2 Finding 2: The "correct" meaning of a $100(1\text{-}\alpha)\%$ confidence interval of an unknown coefficient

Literally, confidence interval is the frequency or probability that includes the true value of the estimated parameter. The confidence level is the frequency (i.e., the proportion) of possible confidence intervals that contain the true value of their corresponding parameter. In other words, if confidence intervals are constructed using a given confidence level in

an infinite number of independent experiments, the proportion of those intervals that contain the true value of the parameter will match the confidence level.

Confidence Interval can be expressed as Equation (4):

$$P\left(L \leq \beta_j \leq U\right) = 1 - \alpha \tag{4}$$

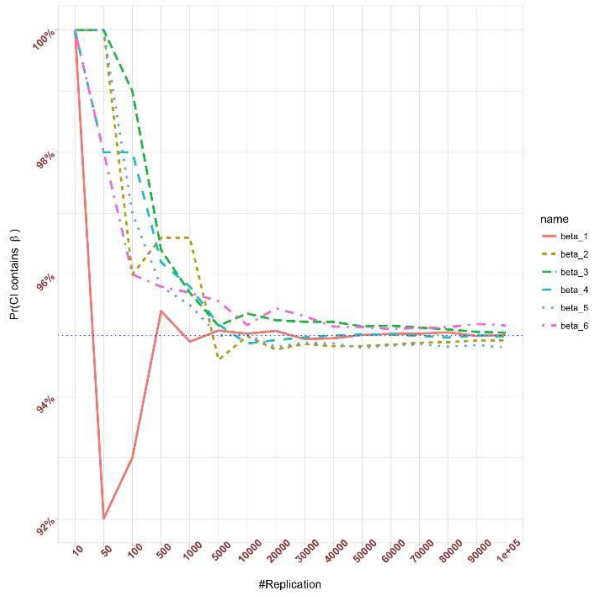

Figure 3. Convergence Graph of Probability of Confidence Interval containing β

The dashed line is the horizon line of 0.95.

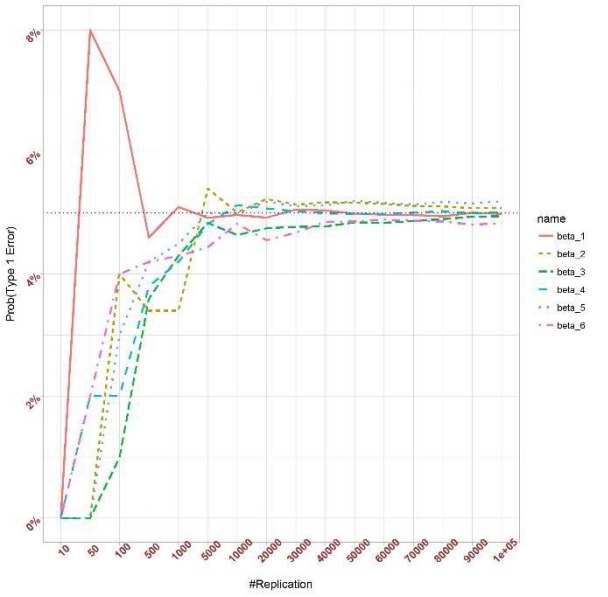

Figure 4. Convergence Graph of Probability of making Type I

The dashed line is horizon line of $\alpha = 0.05$

where consider a coefficient β_j in β, the interval $L \leq \beta_j \leq U$ is a $100(1-\alpha)\%$ confidence interval that will include β_j. In this study, we set α=0.05 earlier for the Fama–French five-factor model. Figure 3 shows confidence intervals of $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4, \hat{\beta}_5$ and $\hat{\beta}_6$ We can see from Figure 4 that with the increase of replication numbers, all the confidence intervals extremely converge to the confidence interval 95% we assigned earlier. Take the variable "HML" for example. If we repeat the sample for 30000 times, roughly 28500 (95%) times the true coefficient value of HML will fall into the generated confidence interval.

3.3 Finding 3: The significance level of the t test for testing a linear hypothesis concerning one or more coefficients is the probability of committing a Type I error

Consider two disjoint parameter space B_0 and B_1, true β is either in B_0 or B_1. We are interested in t test $\psi \in \{0,1\}$ for the null hypothesis $H_0: \hat{\beta} = \beta$ such that: (1) If $\psi = 0$, H_0 is not rejected and $\beta \in B_0$; (2) If $\psi = 1$, H_0 is rejected and $\beta \in B_1$. And we set the significance level $\alpha = 0.05$, then the t-statistics for this t-test is:

$$t = \frac{\hat{\beta}_j - \beta_j}{\sqrt{s^2 \gamma_j}} \tag{5}$$

where s^2 is defined in Equation (3) and γ_j is the j-th diagonal coefficient of $(X^T X)^{-1}$. And t-statistics follows t_{n-k} distribution.

Probability of committing Type I error of the test ψ (rejecting H_0 when it is true) is $\mathbb{P}_{\beta \in B_0}(\psi = 1) = \mathbb{P}_{\beta \in B_0}(t > t_{1-\frac{\alpha}{2}, n-k} \text{ or } t < -t_{1-\frac{\alpha}{2}, n-k})$. According to Table 1, and Figure 4, we can find that the probability is close to $\alpha = 0.05$ when replication number N is sufficiently large. We can say that the significance level of the t test, α, for testing a linear hypothesis concerning coefficients is the probability of committing a Type I error.

Sometimes, Type I error is worse than Type II error, i.e., innocent person goes to jail or bad quality drugs goes to market. From this finding, we can know that if we want to restrict the probability of committing Type I error during null hypothesis testing, we can set confidence level α to an optimal value, say 0.05. Then the probability that we reject the H_0: $\hat{\beta} = \beta$, which means that we embrace the wrong coefficients for the Fama French five-factor model, is at most 5%.

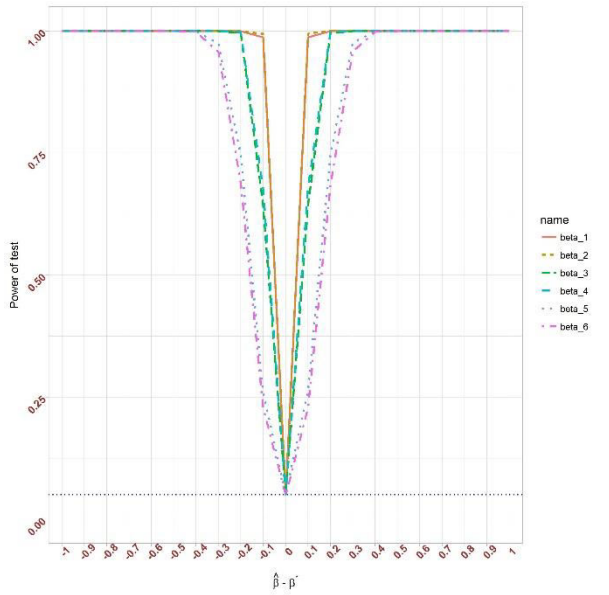

Figure 5. Convergence Graph of Power

The horizon dashed line is the line of 0.05. x-axis label is $\hat{\beta} - \beta^*$.

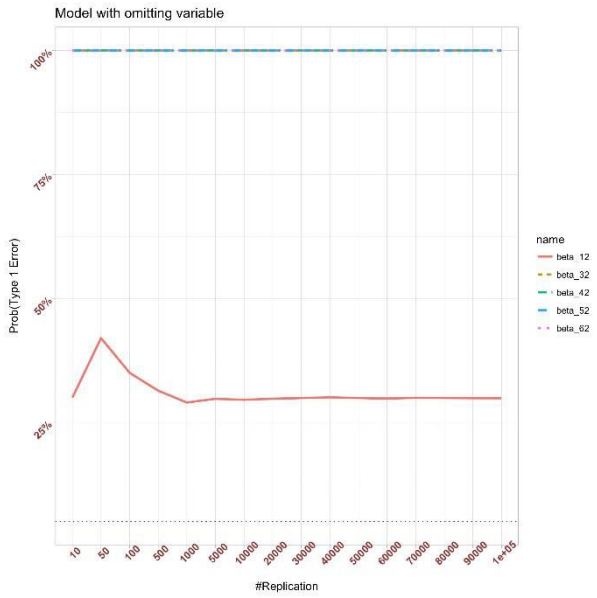

Figure 6. Convergence Graph of Probability of Type I Error in Model with Omitted Variable

We omit variable $R_m - R_f$ here. The horizon dashed line is the line of 0.05

3.4 Finding 4: The t test is unbiased

Probability of committing Type II error of the test ψ (Not rejecting H_0 when it is false) is $\mathbb{P}_{\beta\in B_1}(\psi = 0) = \mathbb{P}_{\beta\in B_1}(-t_{1-\frac{\alpha}{2},n-k} \le t \le t_{1-\frac{\alpha}{2},n-k})$ and power of a test ψ is $\mathbb{P}_{\beta\in B_1}(\psi = 1) = 1 - \mathbb{P}_{\beta\in B_1}(\psi = 0)$. The t test is unbiased if power of test ψ $\mathbb{P}_{\beta\in B_1}(\psi = 1) >$ probability of committing Type I error $\mathbb{P}_{\beta\in B_0}(\psi = 1)$, which means that the probability of rejecting test ψ (power of test ψ) is always higher when the alternative is true than when the null is true. We will build multiple test ψ_k with null hypothesis H_0: $\hat{\beta}_j = \beta_k^*$ and β_k^* ranges from $\beta_j - c$ to $\beta_j + c$ and check the power of t test. According to Figure 5, we can find that power of t test tends to approach 1 as the null hypothesis H_0: $\hat{\beta}_j = \beta_k^*$ becomes more likely untrue, which proves that t test is unbiased.

Through the Figure 5, we can make a consistent conclusion as Section 3.1, which is that $\hat{\beta}_2$, the estimator of the market excess return factor $(R_m - R_f)$ is the most stable factor for the market risk premium, because it has the highest convergence rate to the significance level when $\beta_k^* \to \beta_j$. This can be supported by the variance of OLS estimators, which are listed in Table 1.

3.5 Finding 5: The result in Part iii) no longer holds if some relevant explanatory variables have been omitted from the model

To test whether the result in Section 3.3 still holds or not if explanatory $R_M - R_f$ have been omitted from the model, we redo the simulation of model with omitted variable, and postulate that other coefficients still are equal to their true value. The result shows that with the increase of replication numbers, all convergence values of the probability of committing Type I error are greater than 25% (Figure 6), $\hat{\beta}_{12}, \hat{\beta}_{32}, \hat{\beta}_{42}, \hat{\beta}_{52}, \hat{\beta}_{62}$ are far from their true value (Figure 7),

and the convergence value of error variance is far from the true error variance 1.4060 (Figure 8) after we omit variable $R_M - R_f$.

It means that if some of the important variables are omitted, the model will be more likely to suffer from Type I error. In particular, the null hypothesis of $\hat{\beta}_i = \beta_i$ should be true, but we wrongly reject it. In this situation, the significance level of the t test no longer reflect the probability of committing a Type I error.

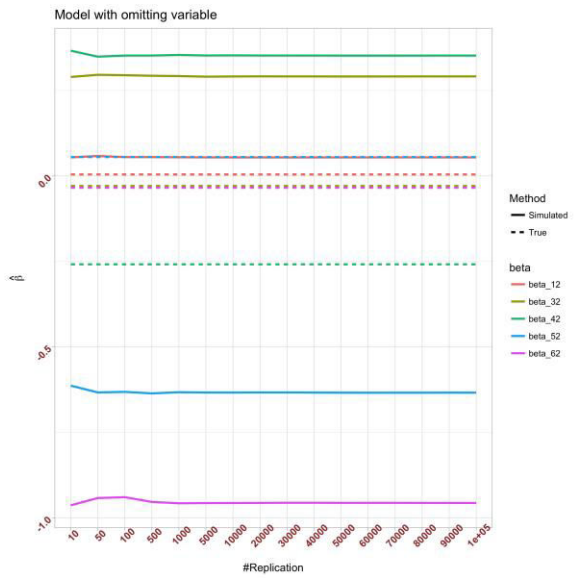

Figure 7. Convergence Graph of $\hat{\beta}_i - \beta_i$ in Model with Omitted Variable

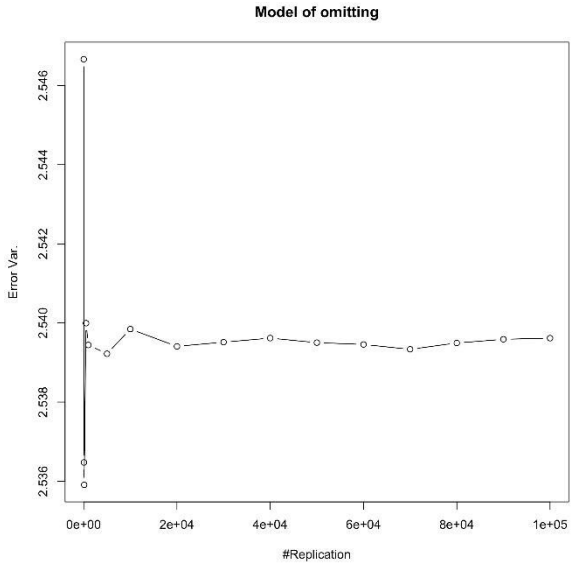

Figure 8. Convergence Graph of Error Variance in Model with Omitted Variable

3.6 Finding 6: The wider implications of "model selection" for statistical modeling and the lessons to be learnt for practitioners

We try to omit different explanatories in each model and redo the regression. The result (Table 2) shows that the AIC and BIC of Model 3, 5, 6, 7 are lower than those of Full Model, and the Fama–French three-factor model is still relatively better than other model combining Adjusted R-squared criterion, Mallow's C_p criterion, and information criterion.

It means that if some unimportant variables (e. g. SMB, RMW, CMA) are include in the model then the variances of coefficient estimators will become inflated. Therefore, Even if we know omitting some of the important variables will cause the estimators of the remaining coefficients become biased, practitioners should keep balance between unbiased estimator and low variance of estimator by using some model selection techniques such as Mallow's C_p.

Table 2. Different Criteria of Each Model

	Adjusted R-squared	AIC	BIC	Mallow's C_p
Full Model	0.4982286	8046.163	8086.981	
Model 2(Omit β_2)	0.1064741	9498.768	9533.756	1963.3724737
Model 3(Omit β_3)	0.4981351	8045.635	8080.622	-0.5302520
Model 4(Omit β_4)	0.4903992	8084.166	8119.153	38.2578132
Model 5(Omit β_5)	0.4981208	8045.706	8080.694	-0.4588785

Model 6(Omit β_6)	0.1064741	8045.325	8080.312	-0.8401955
Model 7(three-factor)	0.4980998	8044.814	8073.970	0.6473720
Model 8(CAPM)	0.4863740	8100.990	8118.484	61.5142284

4 Conclusion

In this study, we firstly do O.L.S to investigate the factors to describe stock returns (i.e. five factors in the Fama–French model). We downloaded data from Jan 1st, 2008 to Dec 29th, 2017 from the Kenneth R. French homepage. Then we got the true corresponding coefficients of each factors by O.L.S regression using real data. After that, we utilize Monte Carlo simulations to test the validity of the Fama–French five-factor model and prove some properties of OLS estimator (such as unbiasedness), confidence interval, t-test, Type I error and model selection.

We confirm the validity of five factors in the Fama–French model. In addition, we compare the relative importance between these factors, and find that CAPM is indeed the most important explanatory, and HML after it. These findings provide a good deal of insight into portfolio management and asset pricing.

Our results also uncovered six interesting theoretical findings through the regression model, 1) The O.L.S. estimators of the unknown coefficients and error variance are unbiased. 2) The "correct" meaning of a $\mathbf{100}(\mathbf{1}\text{-}\boldsymbol{\alpha})\%$ confidence interval of an unknown coefficient is that roughly the true coefficient value of the coefficient will fall into the generated confidence interval in $\mathbf{100}(\mathbf{1}\text{-}\boldsymbol{\alpha})\%$ replications. 3) The significance level of the t test for testing a linear hypothesis concerning one or more coefficients is the probability of committing a Type I error. 4) The t test is unbiased. 5) if some of the important variables are omitted, the estimators of the remaining coefficients will become biased. 6) in model selection, practitioners should keep balance between unbiased estimator and low variance of estimator by using some model selection techniques.

In conclusion, the Monte Carlo simulations help us validate each factor in the Fama–French five-factor model and understand the properties of O.L.S. In the future, we will use O.L.S regression to explore more factors in other asset pricing model.

For detail processing code, please refer to the project github site:
https://github.com/vincent27hugh/FB8916_FF5_MC

Reference

[1] E. F. Fama and K. R. French, "Common risk factors in the returns on stocks and bonds," *Journal of Financial Economics,* vol. 33, pp. 3-56, 1993/02/01/ 1993.

[2] C. Gaunt, "Size and book to market effects and the Fama French three factor asset pricing model: evidence from the Australian stockmarket," *Accounting & Finance,* vol. 44, pp. 27-44, 2004.

[3] A. Brav and A. Gompers Paul, "Myth or Reality? The Long-Run Underperformance of Initial Public Offerings: Evidence from Venture and Nonventure Capital-Backed Companies," *The Journal of Finance,* vol. 52, pp. 1791-1821, 2012.

[4] E. F. Fama and K. R. French, "A five-factor asset pricing model," *Journal of Financial Economics,* vol. 116, pp. 1-22, 2015/04/01/ 2015.

[5] D. P. Kroese, T. Brereton, T. Taimre, and Z. I. Botev, *Why the Monte Carlo method is so important today* vol. 6, 2014.