DATA 201: Time Series Analysis
Linear Regression Model
Lecture 9: Size portfolios

Lulu Wang

Data Analytics
Dickinson College

2/18/2025

Introduction

CAPM and Fama-French

Book-to-Market (BM) portfolios

What is style

Omitted variable bias

Objectives

- Review the Linear Regression Model (LRM)
- Discuss its estimation in Python
- Explore applications in financial data

I will review the regression model in broad terms and more details can be found in an introductory statistics/econometrics textbook, such as:

- Stock and Watson, *Introduction to Econometrics*, Pearson
- Wooldridge, *Introductory Econometrics*, A Modern Approach, South-Western

The Linear Regression Model (LRM)

- The linear regression model is given by:

$$Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t,$$

  where:
  - Y_t: dependent variable at time t
  - X_t: independent variable / factor / predictor at time t
  - β_0, β_1: coefficients to be estimated
  - ε_t: error term (mean zero, variance σ^2)
- The **expected** (or average) value of Y_t given X_t is:

$$E(Y_t \mid X_t) = \beta_0 + \beta_1 X_t.$$

CAPM

- **Interpretation**:
  - β_0: the expected value of Y_t when $X_t = 0$
  - β_1: the expected change of Y_t for a unit change of X_t
- The Capital Asset Pricing Model (CAPM) is an example of an LRM:

$$R_t^i = \beta_0 + \beta_1 R_t^{\mathrm{MKT}} + \varepsilon_t,$$

where R_t^i and R_t^{MKT} represent the excess stock and market returns, respectively.

LRM with Multiple Independent Variables

- Typically, we have several independent variables (factors) relevant to explain the dependent variable.
- We extend the Linear Regression Model (LRM) to include independent variables denoted as $X_{k,t}$ for $k = 1, \ldots, K$:

$$Y_t = \beta_0 + \beta_1 X_{1,t} + \cdots + \beta_K X_{K,t} + \epsilon_t$$

- The model is estimated by OLS, making the formula more complex than the single regressor case.

Fama-French Three-Factor Model

- In asset pricing, independent variables are typically referred to as risk factors.
- Fama and French (1993) extend CAPM with two additional factors:
  - **SMB** (Small-minus-Big): Return spread between small and large capitalization stocks.
  - **HML** (High-minus-Low): Return spread between high and low Book-to-Market ratio stocks (value vs. growth stocks).

FF3 Model Equation

- The Fama-French Three-Factor model is defined as:

$$R_t^i = \beta_0 + \beta_1 MKT_t + \beta_2 SMB_t + \beta_3 HML_t + \epsilon_t$$

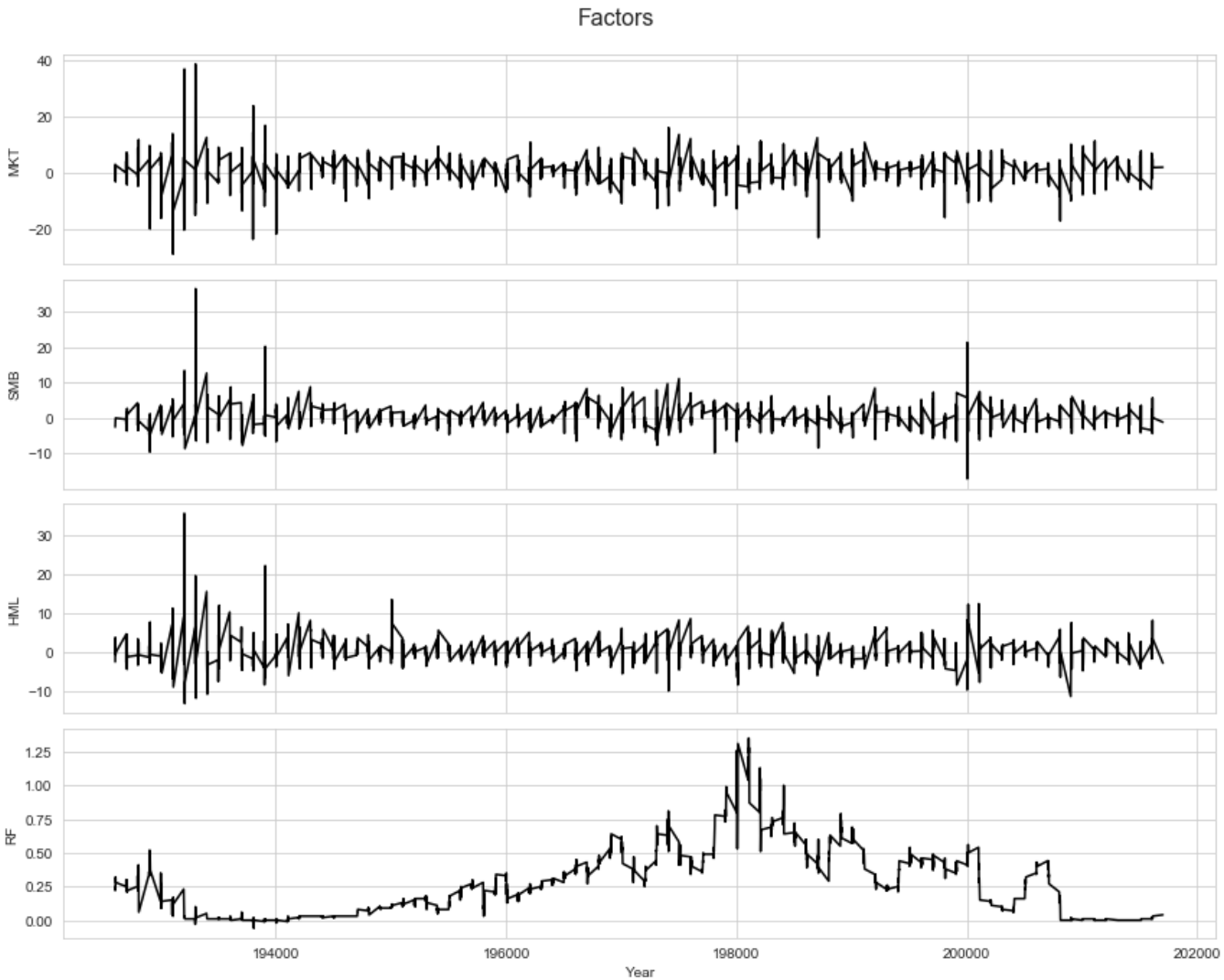
- The goal is to decompose portfolio return R_t^i into exposures to:
  - **Systematic risk**: $\beta_1 MKT_t + \beta_2 SMB_t + \beta_3 HML_t$
  - **Idiosyncratic risk**: ϵ_t
  - **Risk-adjusted return**: β_0

Economic Interpretation of FF3 Factors

- **Market Risk (MKT_t)**: Captures overall market movement.
- **SMB (Size Factor)**:
  - Small firms tend to have higher expected returns than large firms.
  - Related to firm size premium in asset pricing.
- **HML (Value Factor)**:
  - Value stocks (high Book-to-Market) outperform growth stocks (low Book-to-Market).
  - Reflects differences in risk exposure between value and growth stocks.

A first look at FF factors

- The data for the factors are downloadable at Ken French webpage (1926-2017)



Book-to-Market (BM) portfolios

- Another criteria that is used to form portfolios is the Book-to-Market (BM) ratio

$$BM = \frac{\text{book value of equity}}{\text{market cap}}$$

- Sorting stocks by BM ratio identifies two groups of stocks:
  - value stocks have high BM ratio
  - growth stocks have low BM ratio

At the end of June the universe of stocks are sorted by BM ratio and portfolios are formed and kept from July of the current year to June of the following year

Book-to-Market (BM) portfolios (Cont.)

- Value stocks (Hi 10) have an average monthly return of 1.080% vs 0.586% for growth stocks, a premium of 0.494% monthly
- Is it risk?

	AV RET	STD DEV
Lo 10	0.586	5.671
Dec 2	0.704	5.309
Dec 3	0.687	5.406
Dec 4	0.657	5.926
Dec 5	0.738	5.614
Dec 6	0.804	6.048
Dec 7	0.729	6.414
Dec 8	0.925	6.751
Dec 9	1.061	7.626
Hi 10	1.080	9.145

CAPM and BM portfolios

- The high decile BM portfolios (value stocks) earn higher average returns which is (partly) due to higher sensitivity/exposure to market fluctuations.
- The intercept estimates are larger than zero mostly for the value portfolios, but are smaller relative to the size portfolios.

```
Coefficient estimates for the CAPM model on 10 portfolio sorted by book-to-market ratio:

        Lo 10  Dec 2  Dec 3  Dec 4  Dec 5  Dec 6  Dec 7  Dec 8  Dec 9  Hi 10
const -0.083  0.074  0.042 -0.040  0.080  0.113    0.0  0.163  0.218  0.114
MKT    1.009  0.950  0.973  1.052  0.993  1.042    1.1  1.150  1.271  1.457
```

- The t-statistics show that the **alpha** for the 8th and 9th decile portfolios are significant at 5%

Lo 10	Dec 2	Dec 3	Dec 4	Dec 5	Dec 6	Dec 7	Dec 8	Dec 9	Hi 10
-1.554	1.578	0.938	-0.713	1.439	1.575	0.005	1.918	2.068	0.782

3-Factor and BM portfolios

- Findings:
  - Alpha are mostly small and negative and a few small and positive
  - MKT exposure quite close to 1 for all portfolios
  - Small negative exposure to SMB from Lo 10 to Dec 6 and then increasingly positive up to 0.554 for Hi 10

Coefficient estimates of the 3-factor model on 10 portfolio sorted by book-to-market ratio:

	Lo 10	Dec 2	Dec 3	Dec 4	Dec 5	Dec 6	Dec 7	Dec 8	Dec 9	Hi 10
const	0.013	0.129	0.057	-0.080	0.018	-0.001	-0.152	-0.030	-0.034	-0.229
MKT	1.076	0.984	0.987	1.033	0.967	0.984	1.008	1.020	1.090	1.179
SMB	-0.077	-0.008	-0.043	-0.025	-0.068	-0.058	0.021	0.121	0.237	0.554
HML	-0.324	-0.197	-0.040	0.150	0.244	0.428	0.547	0.664	0.844	1.083

  - The 3-factor model improves the fit for all portfolios, but in particular for the high decile portfolios that are not well priced by te CAPM model

Adjusted R-square for the CAPM and the 3-factor models estimate on the Book-to-Market portfolios:

	Lo 10	Dec 2	Dec 3	Dec 4	Dec 5	Dec 6	Dec 7	Dec 8	Dec 9	Hi 10
CAPM	0.904	0.915	0.924	0.901	0.894	0.848	0.839	0.828	0.793	0.725
3-FACTOR	0.945	0.931	0.926	0.908	0.916	0.906	0.924	0.945	0.948	0.929

Presenting the estimation results

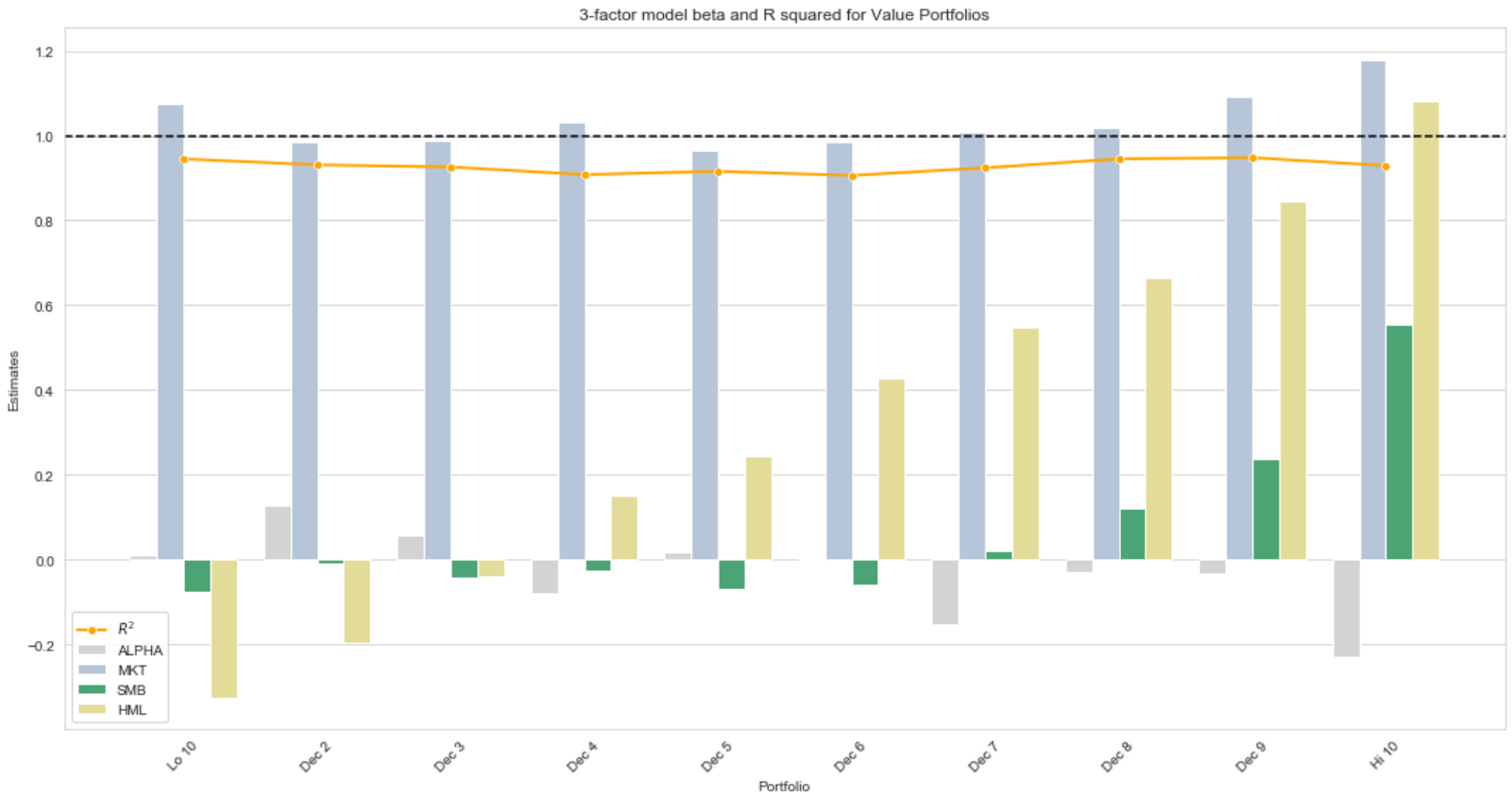
- Presenting the estimation results:

Estimation results for the CAPM model on the 10 value portfolios.

	Alpha	Alpha_t-stat	MKT	MKT_t-stat	SMB	SMB_t-stat	HML	HML_t-stat	Adj_R2
Lo 10	0.0126	0.310	1.0756	131.892	-0.0766	-5.738	-0.3243	-27.328	0.945
Dec 2	0.1293	3.043	0.9835	115.780	-0.0076	-0.548	-0.1974	-15.967	0.931
Dec 3	0.0569	1.263	0.9871	109.615	-0.0429	-2.911	-0.0400	-3.054	0.926
Dec 4	-0.0800	-1.460	1.0328	94.197	-0.0246	-1.373	0.1504	9.427	0.908
Dec 5	0.0182	0.366	0.9666	97.351	-0.0681	-4.189	0.2442	16.899	0.916
Dec 6	-0.0007	-0.012	0.9842	86.941	-0.0578	-3.121	0.4284	26.002	0.906
Dec 7	-0.1524	-2.826	1.0076	93.434	0.0213	1.209	0.5475	34.883	0.924
Dec 8	-0.0298	-0.619	1.0200	105.866	0.1208	7.664	0.6644	47.383	0.945
Dec 9	-0.0335	-0.629	1.0905	102.314	0.2373	13.605	0.8435	54.381	0.948
Hi 10	-0.2289	-3.075	1.1786	79.175	0.5535	22.721	1.0830	49.985	0.929

An alternative way of presenting the estimation results

- An alternative way of presenting the estimation results using a graph



A 3 factor model for DFSVX

- The R^2 of the regression is 0.729. All factors are statistically significant at 1%, but the intercept is not
- The large and positive coefficients for SMB and HML indicate significant exposure to these factors: the fund manager over-weights small caps and value stocks
- Even if we did not know the style of the fund, we could have inferred the strategy from the coefficient estimates

```
=== OLS Results (HAC/ Newey-West Standard Errors) ===
                            OLS Regression Results
==============================================================================
Dep. Variable:                  DFSVX   R-squared:                       0.732
Model:                            OLS   Adj. R-squared:                  0.729
Method:                 Least Squares   F-statistic:                     459.8
Date:                Mon, 17 Feb 2025   Prob (F-statistic):          3.07e-108
Time:                        20:48:42   Log-Likelihood:                -738.20
No. Observations:                 286   AIC:                             1484.
Df Residuals:                     282   BIC:                             1499.
Df Model:                           3
Covariance Type:                  HAC
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.0685      0.141      0.487      0.627      -0.208       0.345
Mkt-RF         0.9668      0.036     27.079      0.000       0.897       1.037
SMB            0.8193      0.071     11.512      0.000       0.680       0.959
HML            0.6079      0.076      8.037      0.000       0.460       0.756
==============================================================================
Omnibus:                       83.397   Durbin-Watson:                   2.872
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1647.674
Skew:                           0.595   Prob(JB):                         0.00
Kurtosis:                      14.698   Cond. No.                         4.64
==============================================================================
```

What Is Style?

- Assuming we 7 mistery mutual funds called FUND1 to FUND7
- Then I estimate the Fama-French 3-factor model and the coefficient estimates are shown in the Table, together with the R^2 of the regression
- Can you guess the investment style (large/small cap, value/growth stocks, and combinations) of each mutual fund?

	FUND1	FUND2	FUND3	FUND4	FUND5	FUND6	FUND7
(Intercept)	-0.152	-0.374	-0.356	-0.510	-0.507	-0.232	-0.128
MKT	0.989	1.024	0.996	1.077	1.092	1.056	0.998
SMB	-0.174	0.699	0.828	0.025	0.773	0.592	-0.186
HML	0.013	0.362	0.428	0.413	0.635	0.183	-0.112
R squared	0.987	0.949	0.938	0.940	0.950	0.950	0.964

Omitted variable bias

- Assume that you are given the returns of another mistery fund without being told the identity and style of the fund
- Your task is to analyze the fund returns and evaluate the risks that it is exposed to
- A possible starting point in the analysis is to consider the MKT, SMB, and HML factors to evaluate if the fund invests in US equities

OLS Regression Results

Dep. Variable:	ODMAX	R-squared:	0.369
Model:	OLS	Adj. R-squared:	0.361
Method:	Least Squares	F-statistic:	54.30
Date:	Mon, 17 Feb 2025	Prob (F-statistic):	8.93e-27
Time:	16:45:08	Log-Likelihood:	-780.58
No. Observations:	242	AIC:	1569.
Df Residuals:	238	BIC:	1583.
Df Model:	3		
Covariance Type:	HAC		

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.5801	0.456	1.271	0.204	-0.315	1.475
Mkt-RF	0.8921	0.143	6.234	0.000	0.612	1.173
SMB	0.2554	0.112	2.287	0.022	0.037	0.474
HML	-0.3315	0.315	-1.051	0.293	-0.949	0.286

Omnibus:	174.753	Durbin-Watson:	2.232
Prob(Omnibus):	0.000	Jarque-Bera (JB):	10026.355
Skew:	2.117	Prob(JB):	0.00
Kurtosis:	34.248	Cond. No.	4.86

Omitted variable bias(Cont.)

- The estimation results indicate that the fund is significantly exposed to MKT and SMB, but it has no exposure to HML
- All good?

OLS Regression Results

Dep. Variable:	ODMAX	R-squared:	0.369
Model:	OLS	Adj. R-squared:	0.361
Method:	Least Squares	F-statistic:	54.30
Date:	Mon, 17 Feb 2025	Prob (F-statistic):	8.93e-27
Time:	16:45:08	Log-Likelihood:	-780.58
No. Observations:	242	AIC:	1569.
Df Residuals:	238	BIC:	1583.
Df Model:	3		
Covariance Type:	HAC		

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.5801	0.456	1.271	0.204	-0.315	1.475
Mkt-RF	0.8921	0.143	6.234	0.000	0.612	1.173
SMB	0.2554	0.112	2.287	0.022	0.037	0.474
HML	-0.3315	0.315	-1.051	0.293	-0.949	0.286

Omnibus:	174.753	Durbin-Watson:	2.232
Prob(Omnibus):	0.000	Jarque-Bera (JB):	10026.355
Skew:	2.117	Prob(JB):	0.00
Kurtosis:	34.248	Cond. No.	4.86

Omitted variable bias(Cont.)

- Not really! Turns out that the fund is Invesco Developing Markets (ticker: ODMAX) and does not invest in US (but emerging markets) equities
- Why are the US-based MKT and SMB factors significant when the fund does not invest in US stocks?

Omitted variable bias(Cont.)

- Because:
  - The fund returns are mostly influenced by the dynamics of emerging equity markets.
  - Emerging equity markets positively co-vary with the US equity market.
  - The estimation results find that the fund returns co-vary with the US factors because they co-vary with the emerging market factor.
  - Since we do not include in the regression the emerging market factor, OLS find commonality of movement of the fund returns with MKT and SMB
- Since we omit the EM factor from the regression, the included factors (MKT and SMB) turn out to be significant because of their correlation with the omitted variable.

Omitted variable bias(Cont.)

- Let's consider the MSCI Emerging Markets (EM) Index as a possible benchmark
- What happens if we include EM as an additional explanatory variable for ODMAX?

OLS Regression Results

Dep. Variable:	ODMAX	R-squared:	0.999
Model:	OLS	Adj. R-squared:	0.999
Method:	Least Squares	F-statistic:	1.054e+05
Date:	Mon, 17 Feb 2025	Prob (F-statistic):	0.00
Time:	16:54:24	Log-Likelihood:	74.906
No. Observations:	242	AIC:	-139.8
Df Residuals:	237	BIC:	-122.4
Df Model:	4		
Covariance Type:	HAC		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.1750	0.016	-10.901	0.000	-0.206	-0.144
EMX	0.9980	0.002	465.068	0.000	0.994	1.002
Mkt-RF	0.0036	0.003	1.042	0.298	-0.003	0.010
SMB	0.0024	0.004	0.594	0.553	-0.005	0.010
HML	-0.0042	0.005	-0.921	0.357	-0.013	0.005

Omnibus:	392.798	Durbin-Watson:	0.067
Prob(Omnibus):	0.000	Jarque-Bera (JB):	28.230
Skew:	-0.493	Prob(JB):	7.41e-07

...

Omitted variable bias(Cont.)

- Omitting a relevant variable from the regression distorts the estimates of the other parameters only when there is correlation between the included and omitted variables; to have bias two conditions have to be satisfied:
  1. the omitted variable is a predictor of the dependent variable
  2. the omitted variable is correlated with one (or more) of the included independent variables
- If either of the two conditions fail, than there is no bias from omitting a variable

Omitted variable bias(Cont.)

- In practice:
  - Sometimes we know that a relevant variable is missing but we might not have observations available.
  - Sometimes we do not know what are the relevant variables and we need to include all possible regressors to make sure to eliminate bias in the estimates (ex: hedge fund returns).