DATA 201: Time Series Analysis
Linear Regression Model
Lecture 6: The role of outliers& LRM with multiple independent variables

Lulu Wang

Data Analytics
Dickinson College

2/10/2025

Introduction

CAPM

The role of outliers

LRM with multiple independent variables

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:

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

Application: Stock Return Risk Analysis

- **Objective**: Explain excess return of a stock (risk premium) by decomposing it into different risk components (factors).
- **Two type of risk**
  - *Systematic Risk*: Market-wide influences (e.g., macroeconomic conditions).
  - *Idiosyncratic Risk*: Firm-specific factors (e.g., company news, management changes).

Interpretation and 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.

Newey–West Example in Python

```
=== OLS Results (Default Standard Errors) ===
                            OLS Regression Results
==============================================================================
Dep. Variable:                ep_crsp   R-squared:                       0.034
Model:                            OLS   Adj. R-squared:                  0.023
Method:                 Least Squares   F-statistic:                     3.061
Date:                Mon, 03 Feb 2025   Prob (F-statistic):             0.0837
Time:                        13:02:36   Log-Likelihood:                -392.57
No. Observations:                  89   AIC:                             789.1
Df Residuals:                      87   BIC:                             794.1
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         -0.1903      5.274     -0.036      0.971     -10.673      10.292
DP             2.1865      1.250      1.749      0.084      -0.298       4.671
==============================================================================
Omnibus:                        3.676   Durbin-Watson:                   1.852
Prob(Omnibus):                  0.159   Jarque-Bera (JB):                3.340
Skew:                          -0.474   Prob(JB):                        0.188
Kurtosis:                       3.021   Cond. No.                         10.9
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

```
=== OLS Results (HAC/ Newey-West Standard Errors) ===
                            OLS Regression Results
==============================================================================
Dep. Variable:                ep_crsp   R-squared:                       0.034
Model:                            OLS   Adj. R-squared:                  0.023
Method:                 Least Squares   F-statistic:                     2.135
Date:                Mon, 03 Feb 2025   Prob (F-statistic):              0.148
Time:                        13:02:41   Log-Likelihood:                -392.57
No. Observations:                  89   AIC:                             789.1
Df Residuals:                      87   BIC:                             794.1
Df Model:                           1
Covariance Type:                  HAC
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -0.1903      5.626     -0.034      0.973     -11.218      10.837
DP             2.1865      1.496      1.461      0.144      -0.746       5.119
==============================================================================
Omnibus:                        3.676   Durbin-Watson:                   1.852
Prob(Omnibus):                  0.159   Jarque-Bera (JB):                3.340
Skew:                          -0.474   Prob(JB):                        0.188
Kurtosis:                       3.021   Cond. No.                         10.9
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity and autocorrelation robust (HAC) using 1 lags
```

- The coefficient estimates $(\hat{\beta}_0, \hat{\beta}_1)$ remain the same.
- Standard errors and p-values change to account for heteroskedasticity and/or autocorrelation.

Understanding Outliers

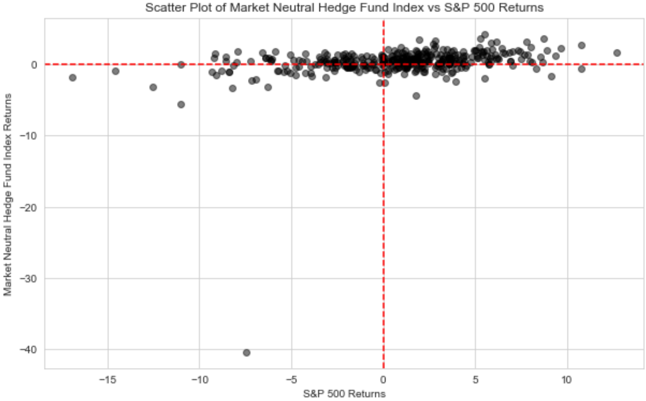
- **Outliers** can be defined as those observations in a sample that are *extreme* relative to most other observations.
- When is an observation considered extreme?
  - Extremes are considered those observations that are 3/4/5 standard deviations away from the sample mean.
- These are extreme observations that arise from:
  - **Exogenous** (e.g., COVID-19).
  - **Endogenous** (e.g., 2008 Global Financial Crisis) to the economic and financial system.

Why Are Outliers Problematic?

- From an **econometric** point of view, outliers are problematic because:
    - They **bias** the coefficient estimates.
    - This leads to estimates deviating from their *true* values.

Example: Equity market neutral HF Index

- The outlier happened in November 2008
- What happened in that month to cause a drop of the EMN Index by 40.45?



Effect of outliers on OLS estimates

- What is the effect of outliers on descriptive statistics? If we dropped the observation for **November 2008**, then:
  - the mean would increase from **0.37%** to **0.48%**
  - the standard deviation would decline from **2.43%** to **1.19%**
  - the skewness would change from **-12.80** to **-0.61**
  - the excess kurtosis would change from **211.87** to **2.72**
- Outliers have large effects on these quantities; what about on the regression estimates?

Effect of outliers on OLS estimates (Cont.)

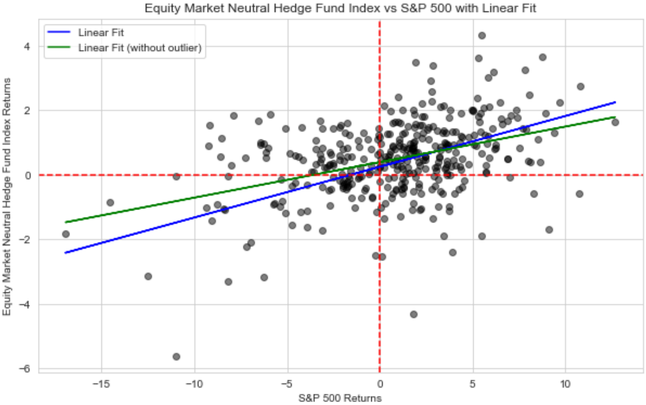
- The regression results show that by dropping the outlier:
  - the slope coefficient decreases from 0.1573 to 0.1101
  - the intercept coefficient increases from 0.2454 to 0.3898

```
=== OLS Results (HAC/ Newey-West Standard Errors) ===
                            OLS Regression Results
==============================================================================================================
Dep. Variable:     Credit Suisse Equity Market Neutral Hedge Fund Index   R-squared:                      0.079
Model:                                                              OLS   Adj. R-squared:                 0.077
Method:                                                   Least Squares   F-statistic:                    10.33
Date:                                                  Mon, 10 Feb 2025   Prob (F-statistic):           0.00142
Time:                                                          13:49:03   Log-Likelihood:               -842.62
No. Observations:                                                   372   AIC:                            1689.
Df Residuals:                                                       370   BIC:                            1697.
Df Model:                                                             1
Covariance Type:                                                    HAC
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.2454      0.154      1.592      0.111      -0.057      0.548
GSPC           0.1573      0.049      3.214      0.001       0.061      0.253
==============================================================================
Omnibus:                      740.617   Durbin-Watson:                   2.001
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           757248.018
Skew:                         -13.126   Prob(JB):                         0.00
Kurtosis:                     222.466   Cond. No.                         4.49
==============================================================================
```

```
=== OLS Results Without Outlier (HAC/ Newey-West Standard Errors) ===
                            OLS Regression Results
==============================================================================================================
Dep. Variable:     Credit Suisse Equity Market Neutral Hedge Fund Index   R-squared:                      0.161
Model:                                                              OLS   Adj. R-squared:                 0.158
Method:                                                   Least Squares   F-statistic:                    43.46
Date:                                                  Mon, 10 Feb 2025   Prob (F-statistic):          1.50e-10
Time:                                                          17:05:33   Log-Likelihood:               -557.74
No. Observations:                                                   371   AIC:                            1119.
Df Residuals:                                                       369   BIC:                            1127.
Df Model:                                                             1
Covariance Type:                                                    HAC
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.3898      0.061      6.394      0.000       0.270      0.509
GSPC           0.1101      0.017      6.592      0.000       0.077      0.143
==============================================================================
Omnibus:                       39.228   Durbin-Watson:                   1.871
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               84.599
Skew:                          -0.566   Prob(JB):                     4.26e-19
Kurtosis:                       5.048   Cond. No.                         4.48
==============================================================================
```

Effect of outliers on OLS estimates (Cont.)

- The scatter plot excluding the extreme observation together with the regression lines estimated with/without the outlier



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.

Multicollinearity in Multiple Regression

- Before estimating the model, analyze the correlation among the $X_{k,t}$ variables:
  - **Perfect collinearity**: If two (or more) independent variables have correlation $= 1$, the LRM cannot be estimated.
  - **Imperfect collinearity**: If two (or more) independent variables have high correlation, estimates become unstable (high variance).

Multicollinearity of variables produces ill-conditioning

Under ordinary polynomials, monomials forming X are highly correlated, OLS coefficients "jump" and the iterative process fails to converge.

Consider an approximation problem $y = Xb + \varepsilon$ such that $y = (0,0)^T$ and

$$\begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 1+\phi & 1 \\ 1 & 1+\phi \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix}.$$

The OLS solution is

$$\hat{b}_1 = \frac{1}{\phi}\left[\frac{\varepsilon_2 - \varepsilon_1(1+\phi)}{2+\phi}\right] \quad \text{and} \quad \hat{b}_2 = \frac{1}{\phi}\left[\frac{\varepsilon_1 - \varepsilon_2(1+\phi)}{2+\phi}\right].$$

Sensitivity of $\hat{b}_1$ and $\hat{b}_2$ to perturbation in $(\varepsilon_1, \varepsilon_2)^T$ is proportional to $1/\phi$.

If $\phi \approx 0$ (multicollinearity), then a small perturbation $(\varepsilon_1, \varepsilon_2)^T$ produces large changes in $\hat{b}_1$ and $\hat{b}_2$.

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:

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

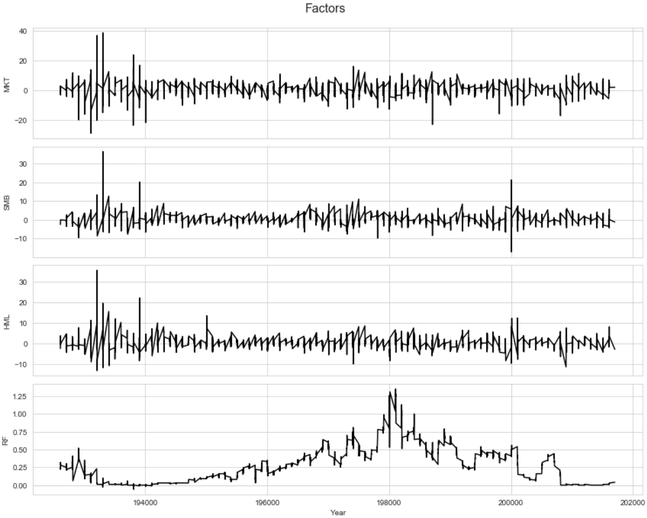
- The goal is to decompose portfolio return R_t^i into exposures to:
  - **Systematic risk**: $\beta_1 MKT_t + \beta_2 SMB_t + \beta_3 HML_t$
  - **Idiosyncratic risk**: ϵ_t
  - **Risk-adjusted return**: β_0

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)



A first look at FF factors(Cont.)

- Compute summary statistics:

	Mean	Std Dev	Skewness	Kurtosis	Max	Min
MKT	0.65	5.37	0.19	7.79	38.85	-29.13
SMB	0.21	3.20	1.89	19.14	36.56	-17.20
HML	0.39	3.52	2.31	20.19	35.61	-13.11
RF	0.28	0.25	1.06	1.25	1.35	-0.06

- Correlation matrix among the Fama-French factors (including the riskfree rate):

	MKT	SMB	HML	RF
MKT	1.00	0.32	0.25	-0.07
SMB	0.32	1.00	0.12	-0.05
HML	0.25	0.12	1.00	0.02
RF	-0.07	-0.05	0.02	1.00