Econometrics and Data Science

Siem Jan Koopman

Topic: **Lessons Learned in Data Science**

School of Business and Economics, VU Amsterdam

Amsterdam Center of Econometrics and Data Science (aceda)

Ksandr Live XL– September 8, 2021

VU VRIJE UNIVERSITEIT AMSTERDAM

aceda

Lessons Learned in Data Science: Elsewhere

8 september 2021

Outline

1 Introduction

2 The Case

3 Mixed, Messy and Noisy Data

4 Econometric Modelling: Causality

5 Conclusion

Econometrics and Data Science: EDS - Department at VU-SBE ACEDA - Data Science Services

EDS Department

- **EDS** stands for

  Department of Econometrics and Data Science
  School of Business and Economics, VU Amsterdam

- **EDS** consists of 30+ fte
  - 15 fte research staff (senior, junior, tenure-track)
  - 3 fte part-time research
  - 3 fte post-docs (external funding)
  - 9 PhD students

ACEDA: Data Science Services

- **Data Science**:
  - Multivariate Statistical Methods
  - Network structures
  - Classification
  - Regression
  - Unsupervised Statistical Learning (k-means, principal components)
  - Random Forests

- **Econometrics**
  - Time Series and Dynamic Econometrics
  - Prediction and Forecasting
  - Causal Structural Modelling
  - Mixed-Frequency, Messy and Noisy Data

Marketing: Correlation and Causality

ACEDA: The Case

- **Marketing Consultancy**:
  - Large Companies, Huge Data Sets
  - Huge Number of Products
  - Many Divisions (and Bosses!)
  - Organized along Different Lines
  - Countries, Products, Channels
  - History of Mergers
  - Communication is Surprisingly Poor

- **Doing Data Science**
  - Challenge in Getting the Data
  - Much Data is Hidden in Systems and Locked
  - Promises Promises in Unlocking
  - Noisy Data with Many Messy Features

Mixed, Messy and Noisy Data

Signal and Noise

Key challenge: from noisy indicators towards SIGNAL EXTRACTION

We focus on **SIGNAL EXTRACTION**:

- Building predictive models from data with
- different frequencies, different features, missing entries, outliers, etc.
- Using Data-Driven frameworks:

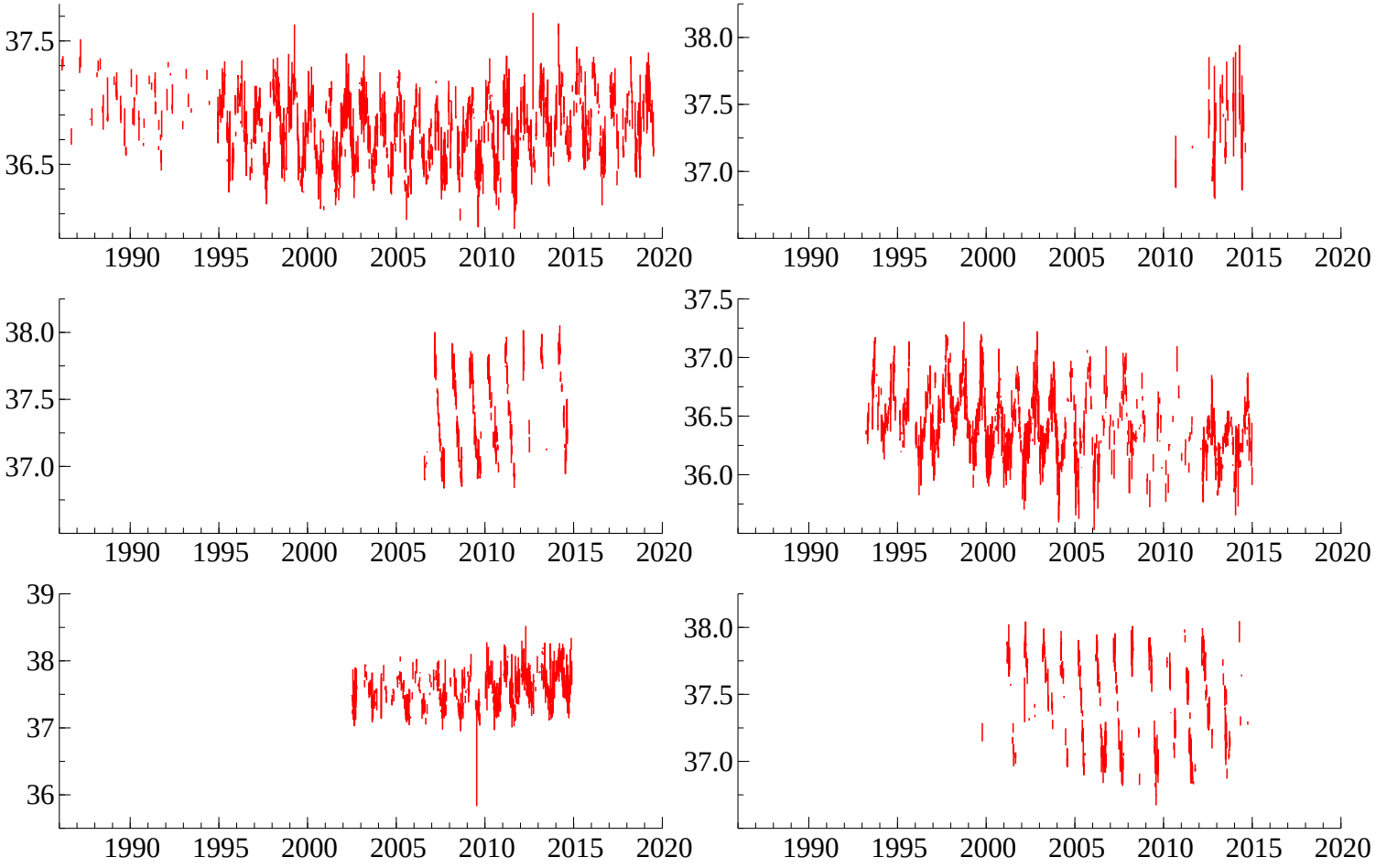
$$y_{it} \sim p_i(y_{it}|\mathcal{F}_{t-1}, \theta_t; \psi), \qquad \theta_t = Z_t \alpha_t$$

- with data y_{it}, model / distribution $p(y_{it}; \cdot)$, past data $\mathcal{F}_{t-1}$, signal θ_t, fixed parameter vector ψ and time-varying parameter vector α_t with dynamic updating

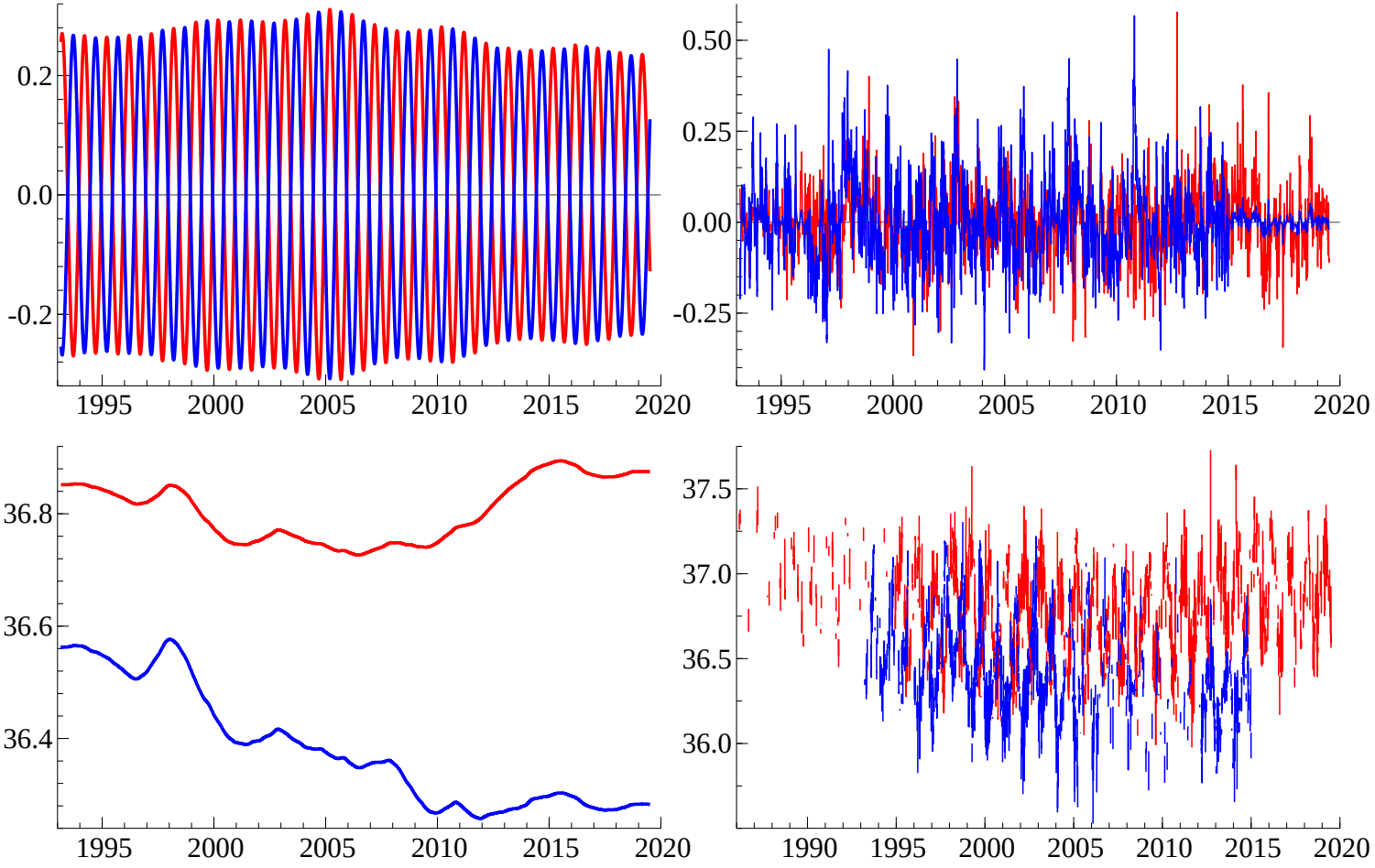
$$\alpha_{t+1} = \omega + \beta \alpha_t + \delta \nabla_t$$

where ∇_t is the innovation (score function) and ω, β, δ are coefficients.

Noisy Data Features



Noisy Data Features

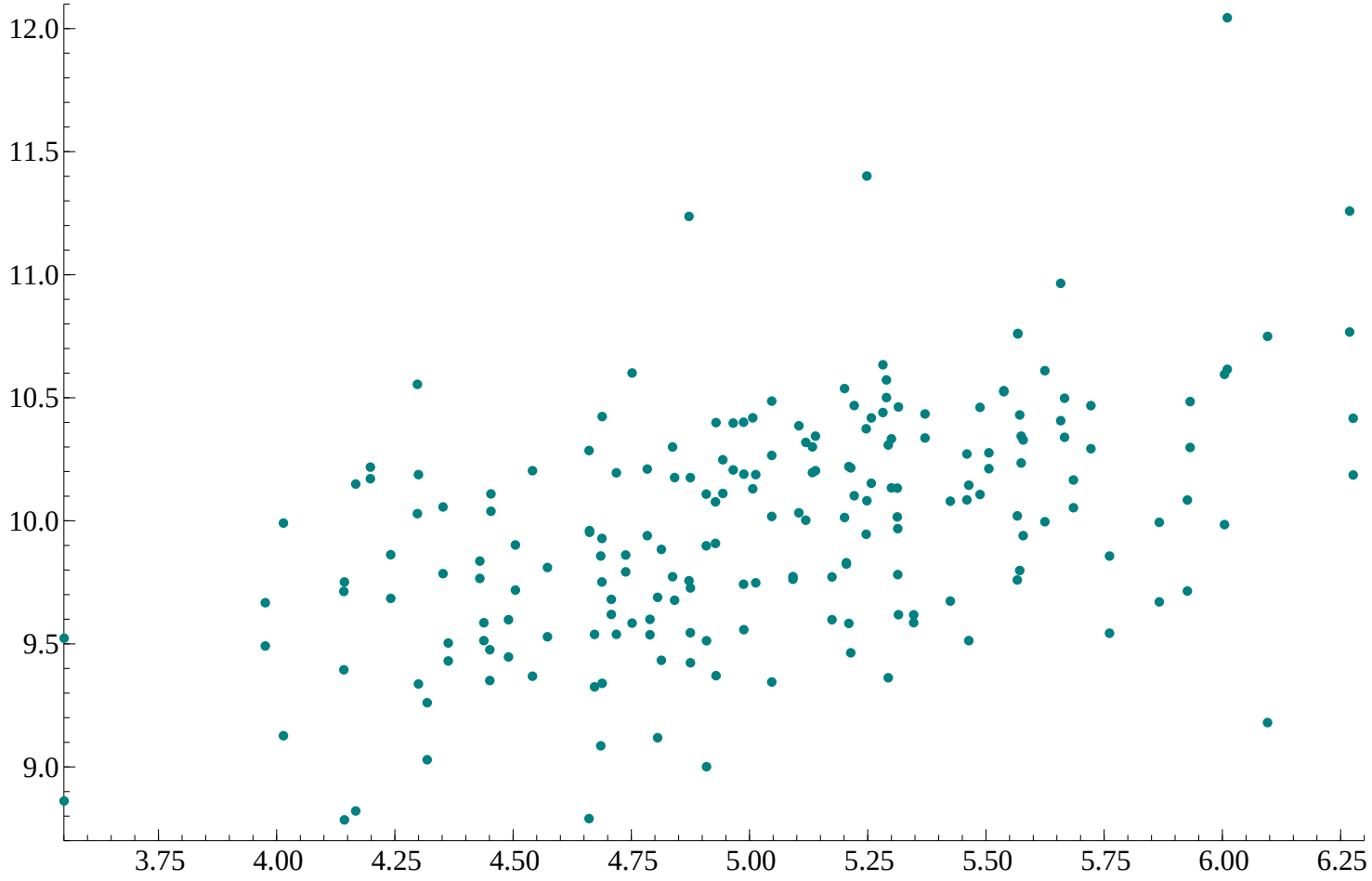


Econometric Causal Modeling

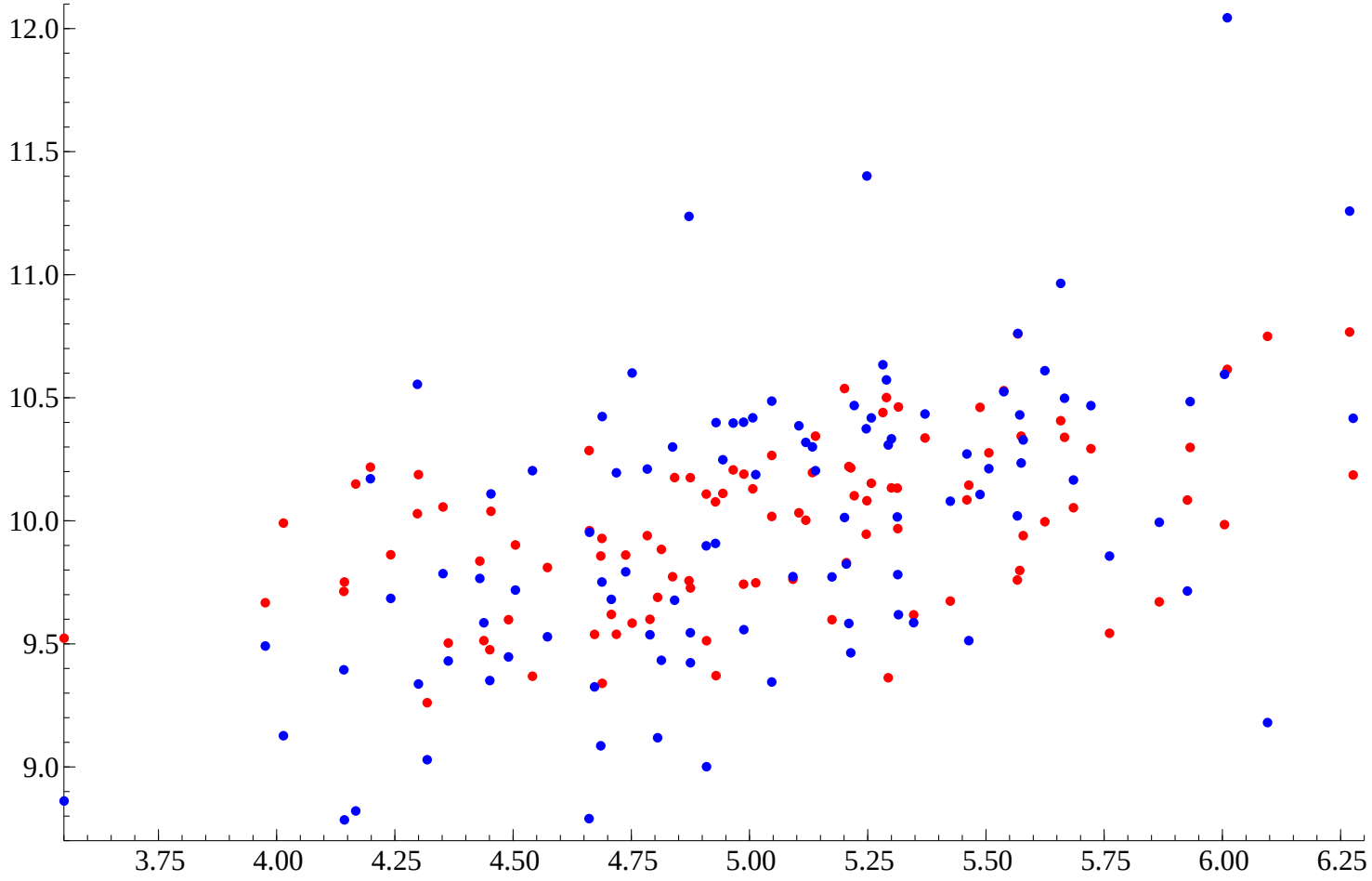
Econometrics and Causality

- Data Science Methods are highly effective in Data Exploration, Visualisation, Correlations
- Correlation versus Causality: structural econometric models
- Classic example is sales/consumption versus price/inflation

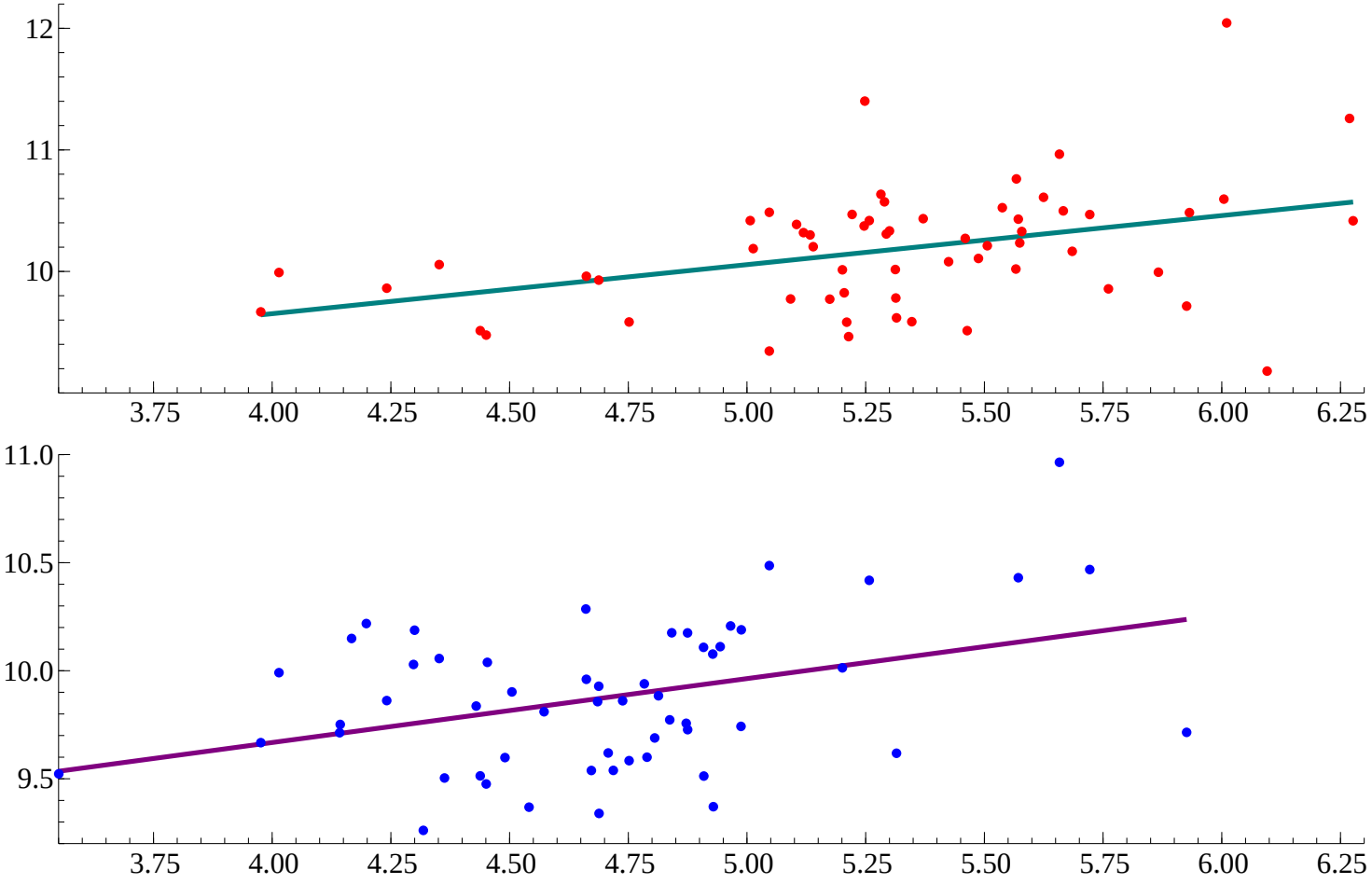
Sales and Discounts



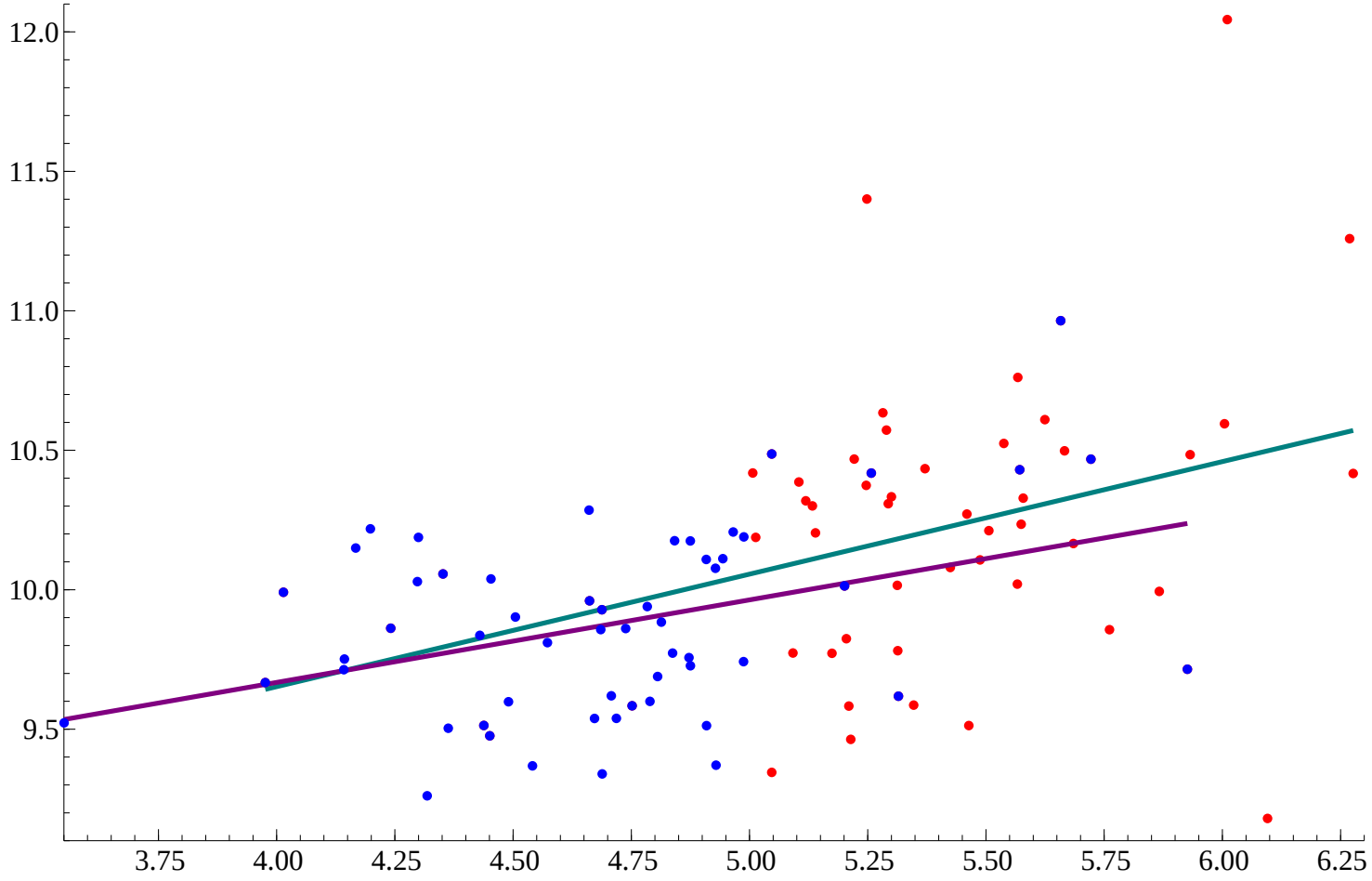
Sales and Discounts: from TWO Divisions



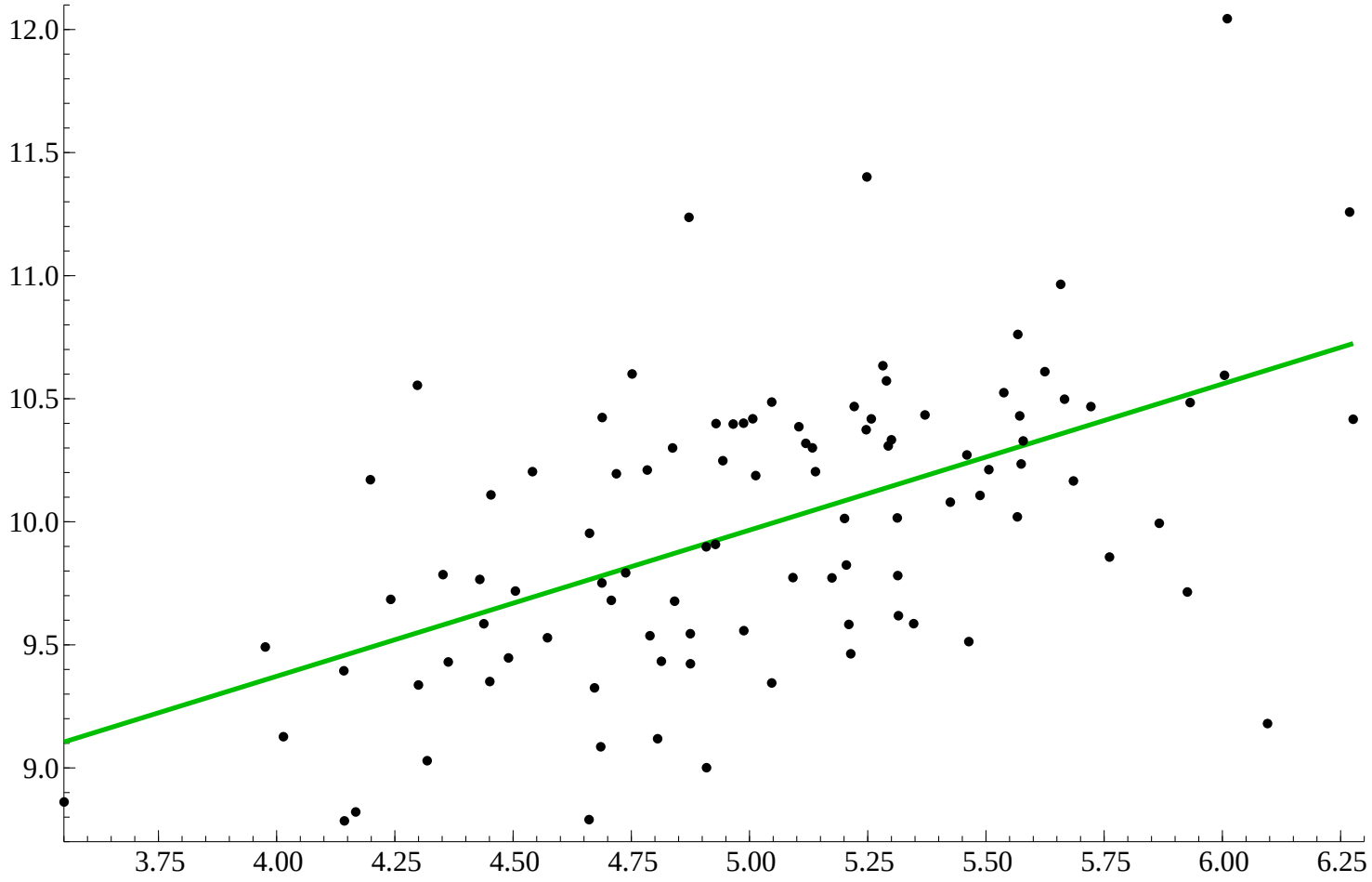
Separate Analysis



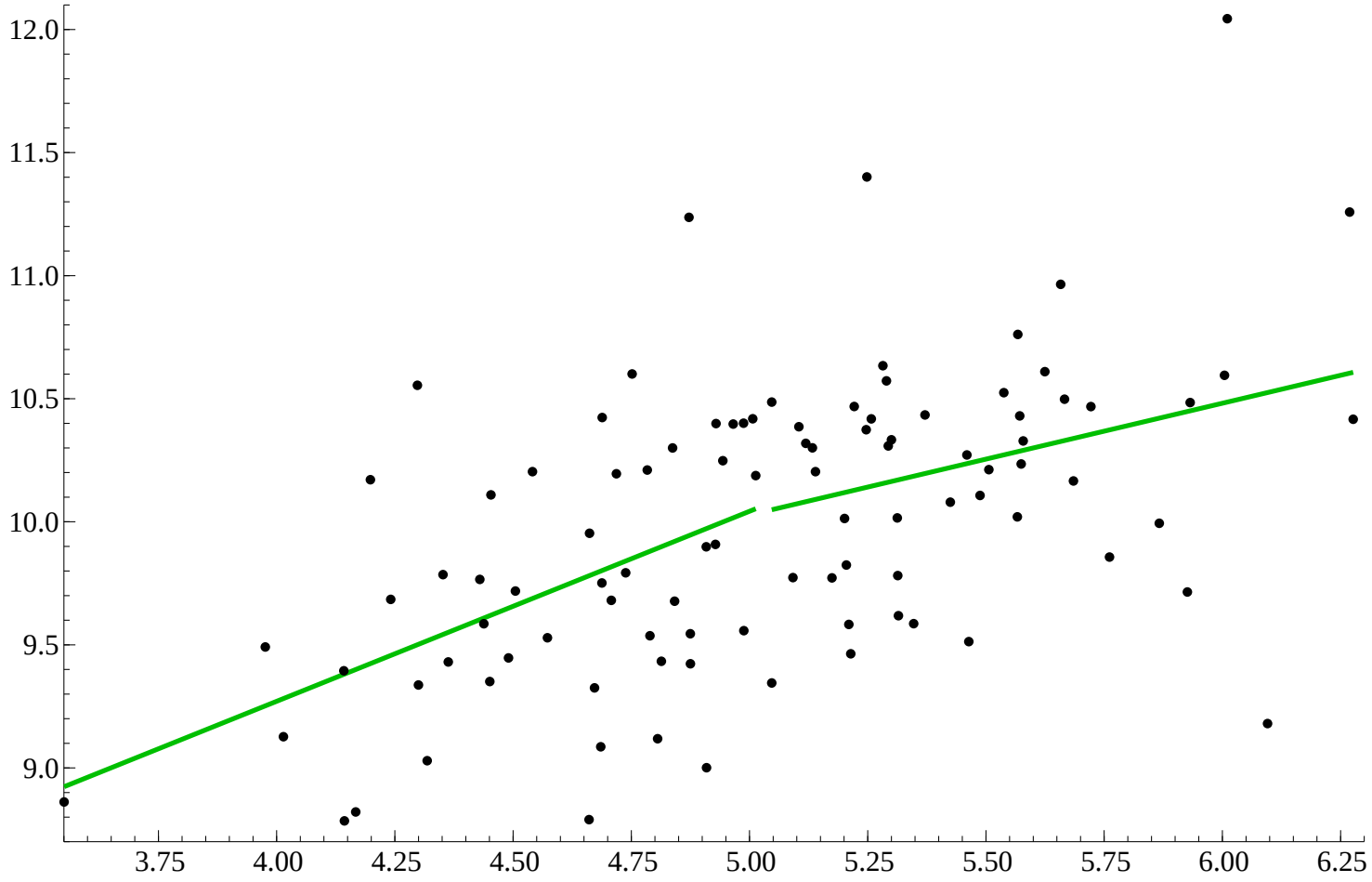
Separate Analysis



Joint Analysis



Joint Analysis



Econometric Analysis

- First: Separate Analyses are Done for the TWO Divisions
- Second: Simultaneous Analyses are Done
- Results are Different
- Relevant Fit Improvements and Policy Ramifications

Conclusion

- EDS and ACEDA
- Broad scope on Econometrics and Data Science Methods
- Structural view on Data Analysis, Modelling and Prediction