JOHNS HOPKINS
CAREY BUSINESS SCHOOL

BU.610.740: Forecasting Models for Business Intelligence
Chapter 2: Time Series Regression

Ali Eshragh

Spring II, 2024

Outline

Linear Models Essentials

- **Linear models** serve as a **fundamental tool** for understanding trends, cycles, and seasonal variations in business data.
- Their **simplicity** and **interpretability** make linear models a key tool for business analysts, **balancing** model complexity with ease of understanding.
- Linear models' applications in **time series** analysis are as extensive and influential as in **classical statistical** data analysis.

Linear Models in Time Series Analysis

- We will express the time series X_t as a **linear combination** of **previous** values of the time series, that is

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + W_t$$

where W_t is a **Gaussian white noise** series.

- The time series X_t may depend on both the **current** and **lagged** values of **another time series**, modeled as follows:

$$X_t = \theta_0 z_t + \theta_1 z_{t-1} + \cdots + \theta_q z_{t-q} + W_t$$

- The time series X_t may also be modeled based on a linear combination of d **predictor time series**, as shown below:

$$X_t = \beta_1 z_{1,t} + \beta_2 z_{2,t} + \cdots + \beta_d z_{d,t} + W_t$$

Simple Linear Regression: Model

- Consider a **simple linear regression** model

$$X_t = \beta_0 + \beta_1 z_t + W_t$$

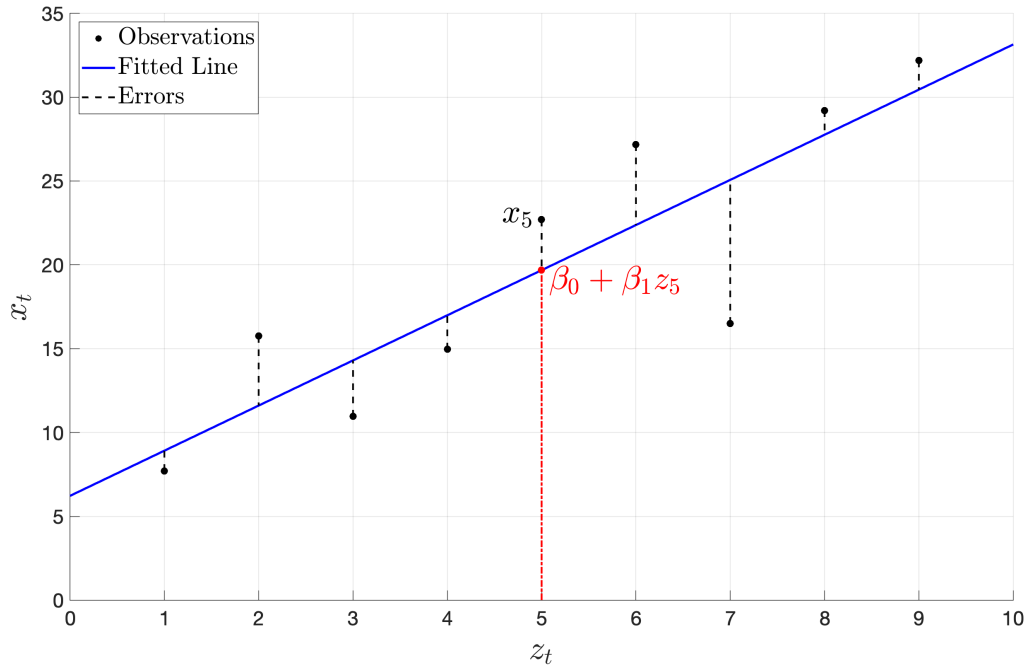
- In a clothing department store, predicting the quarterly **sales revenue** of a product based on the **advertising expenditure** on the product.

- In a retail business online store, predicting the monthly **demand** for a product in terms of the monthly **number of clicks** on the product.

- In an energy provider company, predicting the weekly **electricity consumption** of a building in terms of the average weekly **outdoor temperature**.

Simple Linear Regression: Parameter Estimation

- To **estimate** the unknown coefficients β_0 and β_1, the total **sum of squared errors** is **minimized**: $\text{SSE} = \sum_{t=1}^{n}(x_t - (\beta_0 + \beta_1 z_t))^2$



- The **least squares estimators** for β_0 and β_1 are obtained as follows:

$$\hat{\beta}_1 = \frac{\sum_{t=1}^{n}(z_t - \bar{z})(x_t - \bar{x})}{\sum_{t=1}^{n}(z_t - \bar{z})^2} \quad \text{and} \quad \hat{\beta}_0 = \bar{x} - \hat{\beta}_1 \bar{z}$$

Multiple Linear Regression

- Consider a **multiple linear regression** model

$$X_t = \beta_1 z_{1,t} + \beta_2 z_{2,t} + \cdots + \beta_d z_{d,t} + W_t$$

- In an electronics retailer store, predicting the monthly **product sales** based on monthly **advertising expenditure**, monthly **store visits**, and monthly **customer reviews and ratings** related to the product.

- The unknown coefficients are **estimated** by **minimizing** SSE:

$$\text{SSE} = \sum_{t=1}^{n} (x_t - (\beta_1 z_{1,t} + \beta_2 z_{2,t} + \cdots + \beta_d z_{d,t}))^2$$

Data Representation

- The $n \times d$ **data matrix** $\boldsymbol{Z}$ comprises n samples of d **predictors** at times $t_1, t_2, \ldots, t_n$:

$$\boldsymbol{Z} = \begin{pmatrix} z_{1,1} & z_{2,1} & \cdots & z_{d,1} \\ z_{1,2} & z_{2,2} & \cdots & z_{d,2} \\ \vdots & \vdots & \ddots & \vdots \\ z_{1,n} & z_{2,n} & \cdots & z_{d,n} \end{pmatrix}$$

- The $n \times 1$ **vectors** $\boldsymbol{x}$ and $\boldsymbol{w}$ represent the observed **dependent variable** and **white noise series**:

$$\boldsymbol{x} = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix}^\intercal \quad \text{and} \quad \boldsymbol{w} = \begin{bmatrix} w_1 & w_2 & \cdots & w_n \end{bmatrix}^\intercal$$

- The $d \times 1$ **parameter vector**:

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_1 & \beta_2 & \cdots & \beta_d \end{bmatrix}^\intercal$$

Matrix Representation of the Model

- The **regression** model:

$$\boldsymbol{x} = \boldsymbol{Z}\boldsymbol{\beta} + \boldsymbol{w}$$

- The **sum of squared errors**:

$$\text{SSE} = (\boldsymbol{x} - \boldsymbol{Z}\boldsymbol{\beta})^{\mathsf{T}}(\boldsymbol{x} - \boldsymbol{Z}\boldsymbol{\beta})$$

- The **least squares estimators**, assuming that $Z^{\mathsf{T}}Z$ is invertible and $n > d$:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{Z})^{-1}\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{x}$$

Complexity and Challenges in Big Data Problems

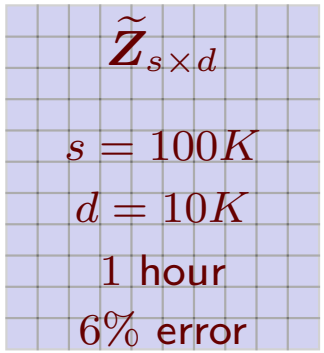
- The **computational time complexity** for least squares estimates is $\mathcal{O}(nd^2)$
- For instance, finding least squares estimators for data with $10,000,000$ **observations** and $10,000$ **predictors** on a 3GHz **CPU** could take 4 days!
- In **big data** regimes with $n \gg d$, $\mathcal{O}(nd^2)$ is **too much**.
- **Machine learning** has developed several approaches in addressing big data challenges, including **randomized numerical linear algebra** (**RandNLA**).
- RandNLA utilizes both **data-oblivious** and **data-aware** random **sub-sampling** methods to **compress** the data matrix Z into a much **smaller** matrix $\widetilde{Z}$, while **preserving** key properties.

Data Matrix Compression

Original Data Matrix:

$s \ll n$

Compressed Data Matrix:



Further Reading

- P. Drineas and M.W. Mahoney, *Lectures on Randomized Numerical Linear Algebra*, arXiv preprint arXiv:1712.08880, 2017
- R. Murray et al., *Randomized Numerical Linear Algebra: A Perspective on the Field With an Eye to Software*, arXiv preprint arXiv:2302.11474, 2023

Assignment 3

Question 1

Identify a problem within **machine learning**, **statistical analysis**, or **optimization** that faces challenges in **big data** regimes. Concisely **define** this problem, **outline** how the challenge related to big data has been tackled, and **cite** at least one relevant reference with a **hyperlink** to the source.

Assignment 3

Example (Expected Answer)

- **Problem:** The least squares problem in statistical regression modeling aims to minimize the sum of squared differences between observed and predicted values, thereby finding the best-fitting linear relationship between the dependent variable and predictors. This optimization technique estimates the model's parameters, ensuring the closest match to the data.
- **Big Data Challenges:** Although efficient algorithms exist to solve the least squares problem, it faces challenges with big data, including computational and storage complexity. Randomized Numerical Linear Algebra is a novel approach to tackle the challenges of least squares problems in big data, using random subsampling algorithms to compress involved matrices and solve a much smaller problem instead, while guaranteeing error bounds on approximation.
- **Refrences:**

[1] P. Drineas, M.W. Mahoney, S. Muthukrishnan, and T. Sarlos, Faster least squares approximation, *Numerische Mathematik*, 117(2):219–249, 2011.

[2] G. Raskutti and M.W. Mahoney, A statistical perspective on randomized sketching for ordinary least-squares, *Journal of Machine Learning Research*, 17(1):7508–7538, 2016.

Model Selection in Regression Modeling

Definition (Model Selection)

Model selection is the process of choosing the **most appropriate** regression model from a set of **candidate** models.

- Linear regression models possess two important **characteristics** in practice: **prediction accuracy** and **interpretability**.
- While overly **complex** models may **excel** in accuracy, they are **harder** to interpret, whereas overly **simple** models **sacrifice** accuracy for clarity.
- There is a **trade-off** between accuracy and interpretability.
- The **goal** is to select the **best minimal subset** of predictors.

General Approaches to Model Selection

- **Forward selection**: Start with an empty model and add predictors one by one based on some statistical performance metrics.
- **Backward elimination**: Start with a full model and remove predictors one by one on some statistical performance metrics.
- **Cross-validation**: Divide the data into training and validation sets, iteratively train models on training set, and select the model with the best performance on the validation set.
- **Regularization methods**: Penalize the model coefficients to prevent overfitting and select the model with optimal regularization parameters such as LASSO and Ridge.

Statistical Performance Metrics: Adjusted R^2

- R-squared **measures** the proportion of the **variation** in the dependent variable that is **explained** by the predictors in a regression model:

$$R^2 := 1 - \frac{\text{SSE}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- Adding **more** predictors to the regression model always leads to a **decrease** in SSE, resulting in an **increase** in R^2

- The Adjusted R^2, a **modified** version of R^2, **addresses** this problem by **penalizing** model complexity based on the number of predictors d:

$$\text{Adjusted } R^2 := \frac{(n-1)R^2 - d}{n - d - 1}$$

- **Higher** Adjusted R^2 values indicate **better-fitting** models.

Statistical Performance Metrics: AIC

- **Akaike Information Criterion** (AIC) measures the expected Kullback-Leibler **distance** between the **true** model and the **fitted** model:

$$\text{AIC} := \log\left(\frac{\text{SSE}}{n}\right) + \frac{2d}{n}$$

- AIC balances the trade-off between **goodness of fit** and model **complexity**.
- **Lower** AIC values indicate **better-fitting** models.
- **Bayesian Information Criterion** (BIC) is similar to AIC but imposes a stronger **penalty** for model complexity:

$$\text{BIC} := \log\left(\frac{\text{SSE}}{n}\right) + \frac{d\log(n)}{n}$$

- Simulation studies suggest that BIC **excels** in large samples, while AIC **outperforms** in smaller samples with many parameters.

Further Reading

Nexus Between Operations Research and Statistics

- D. Bertsimas, A. King, and R. Mazumder, Best subset selection via a modern optimization lens, *Annals of Statistics*, 44(2):813-852, 2016

Multi-objective Optimization Approach

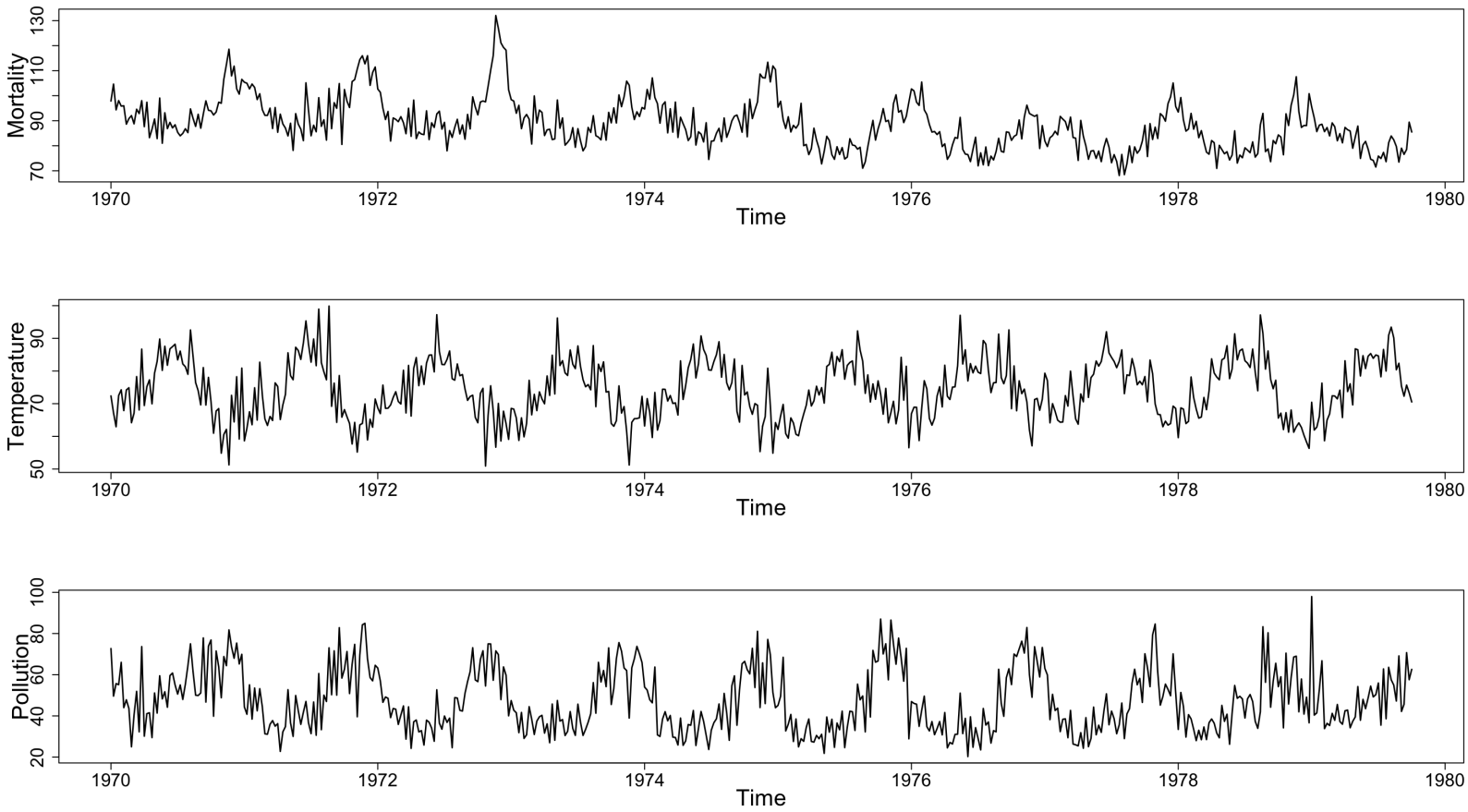
- H. Charkhgard and A. Eshragh, A new approach to select the best subset of predictors in linear regression modeling: Bi-objective mixed integer linear programming, *ANZIAM Journal*, 62(1):64-75, 2019

Overparameterized Deep Learning Models

- L. Hodgkinson, C. van der Heide, R. Salomone, F. Roosta, and M.W. Mahoney, A PAC-Bayesian perspective on the interpolating information criterion, arXiv preprint arXiv:2311.07013, 2023

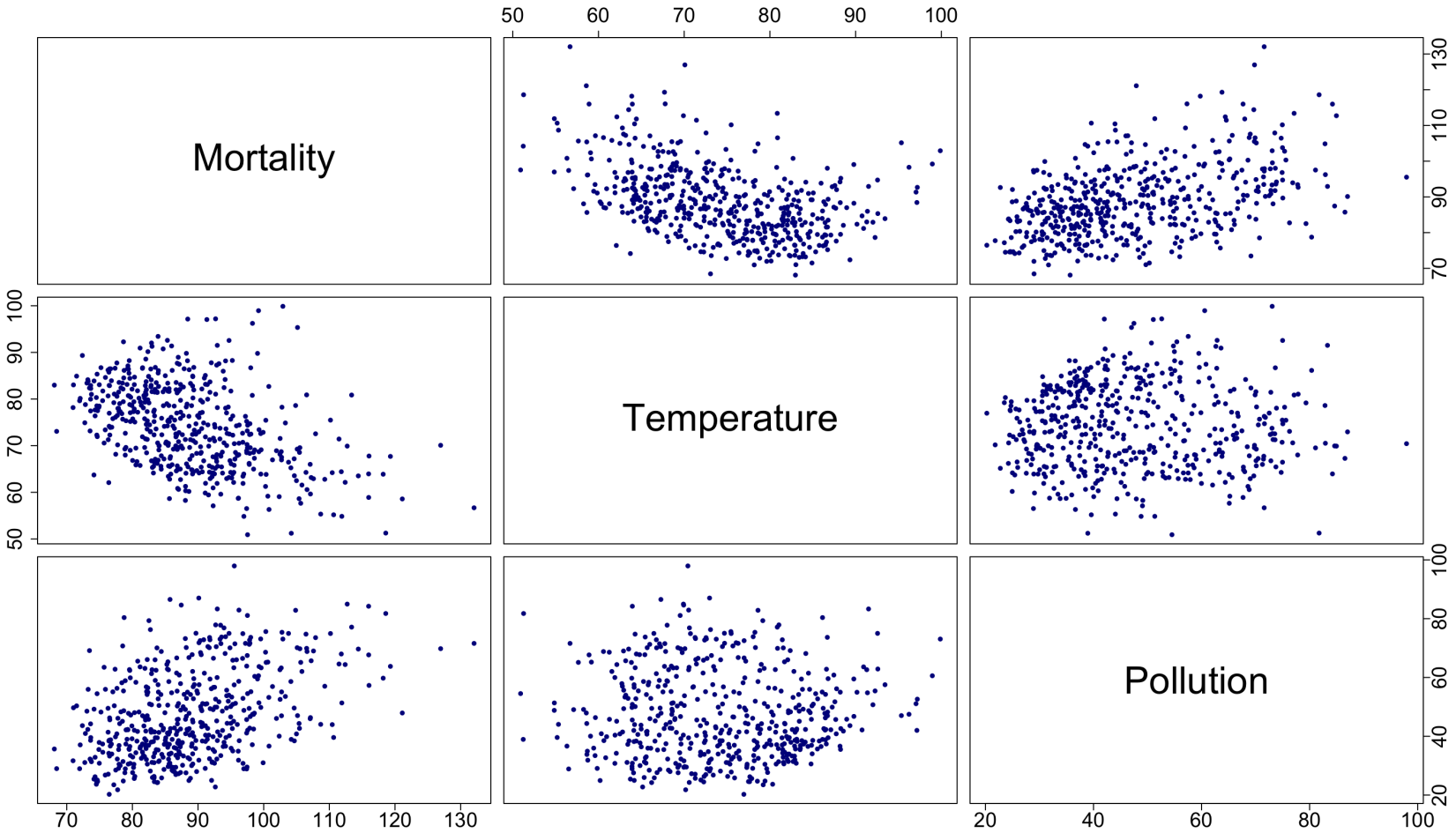
Case Study: Mortality, Temperature, and Pollution

- Investigating the possible effects of **temperature** and **pollution** on weekly **cardiovascular mortality** in Los Angeles County



Case Study: Mortality, Temperature, and Pollution

- Pairwise **scatter plots**



Case Study: Mortality, Temperature, and Pollution

- Let X_t denote **cardiovascular mortality**, y_t denote **temperature**, and z_t denote the **pollution levels** at time t:

  Model 1. $X_t = \beta_{1,1} + \beta_{1,2}t + W_{1,t}$

  Model 2. $X_t = \beta_{2,1} + \beta_{2,2}t + \beta_{2,3}y_t + W_{2,t}$

  Model 3. $X_t = \beta_{3,1} + \beta_{3,2}t + \beta_{3,3}y_t + \beta_{3,4}y_t^2 + W_{3,t}$

  Model 4. $X_t = \beta_{4,1} + \beta_{4,2}t + \beta_{4,3}y_t + \beta_{4,4}y_t^2 + \beta_{4,5}z_t + W_{4,t}$

- **Model selection**:

Model	d	Adjusted R^2	AIC	BIC
1	2	0.21	4.38	4.40
2	3	0.38	4.14	4.17
3	4	0.45	4.03	4.07
4	5	0.59	3.73	3.77

Example

Example (Sales Predictive Analysis)

- An electronics retailer studies how monthly **marketing spend** (m_t), **new products** (p_t), and **customer satisfaction** (r_t) affect **sales** (S_t) through three models:

Model 1. $S_t = \beta_{1,1} + \beta_{1,2} m_t + W_{1,t}$

Model 2. $S_t = \beta_{2,1} + \beta_{2,2} m_t + \beta_{2,3} p_t + W_{2,t}$

Model 3. $S_t = \beta_{3,1} + \beta_{3,2} m_t + \beta_{3,3} p_t + \beta_{3,4} r_t + W_{3,t}$

- **Model selection**:

Model	d	Adjusted R^2	AIC	BIC
1	2	0.35	3.26	3.33
2	3	0.98	0.02	0.13
3	4	0.96	0.06	0.21

References

1. R.H. Shumway and D.S. Stoffer, *Time Series Analysis and Its Applications With R Examples*, Springer, New York, 2017.