Data Analysis for Decision Making

Biwei Chen

Part A

I. Data Structure and Management	II. Data Description and Visualization
III. Statistical Inference and Hypothesis Testing	IV. Regression Models and Analysis

Part B

I. Machine Learning: Regression	II. Machine Learning: Classification
III. Causal Analysis and Inference	IV: Program Impact Evaluation

Lecture 1: Data Structure and Management

Biwei Chen

I. Fundamentals of Data (CH1)

1. What is data? Information or statistics about facts, observations, phenomena
2. Data category: cross-sectional (identities), time series, panel (longitudinal), spatial (geographical)
3. Data is stored in data table(s). A dataset may consist of a single data table or multiple related ones.
4. Data table elements: observations (with IDs) in rows and variables (features) in columns
5. Data quality: content, validity, reliability, comparability, coverage, unbiased selection
6. Data collection and source: API, administration, survey, real-time business, social media
7. Data sampling: 1) population vs samples; 2) random sampling can reduce bias and get to the target
8. Big data three features: large size, high dimension, and complex structure (Goldstein et. al., 2021)
9. Big data applications: automation, machine learning, cloud computing
10. Ethical and legal issues: confidentiality, privacy, sensitivity, uncertainty

II. Data Measurements and Managements (CH2)

1. Data analysts start with structuring and cleaning the data, then turn to data description and analysis.
2. Type of variables: qualitative (factor and category in string) and quantitative (numeric and binary)
3. Type of measurement: flow variables (recorded in a period of time) vs stock variables (a snapshot)
4. Multi-dimensional data presentation: long format (slow vs fast index) and wide format (horizontal)
5. Handling missing data points: 1) deletion; 2) imputation; 3) flag (NA)
6. Relational data: link (combine/join/merge/match) variables from multiple data tables with features
7. Data project organization: 1) raw data tables; 2) clean and tidy data tables; 3) work files for analysis

CH01A Finding a good deal among hotels: data collection
CH01B Comparing online and offline prices: data collection
CH01C Management quality: data collection

CH02A Finding a good deal among hotels: data preparation
CH02B Displaying immunization rates across countries
CH02C Identifying successful football managers

Reference

201705 The Economist, The world's most valuable resource is no longer oil, but data
https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data

202110 IMF Blog, Needed—A Global Approach to Data in the Digital Age
https://blogs.imf.org/2021/10/26/needed-a-global-approach-to-data-in-the-digital-age

Itay Goldstein, Chester S Spatt, Mao Ye, Big Data in Finance, The Review of Financial Studies, 34-7, July 2021, Pages 3213–3225, https://doi.org/10.1093/rfs/hhab038

Lecture 2: Data Exploration and Comparison

I. Exploratory and Descriptive Data Analysis (CH3)

1. EDA or DDA describes the features of the variables of interest in the data
2. Frequencies, probabilities and distributions (all detailed information)
   1) Probability density function PDF: P(X) and cumulative probability function CDF: P(X<=x)
   2) Probability distribution f(X) and cumulative probability distribution F(X)
3. Summary statistics: central values, spread, skewness, kurtosis
   1) Central value: mean, median, mode, quantiles, and percentiles
   2) Spread: range, min, max, variance and standard deviation
   3) Skewness (mean-median difference) (tail thickness) and Kurtosis (peak)
   4) Moments: mean, variance, skewness, and kurtosis
4. Visualization: histogram (discrete), density plot (continuous), box plot, and violin plot
5. Guidelines for data visualization: purpose, focus, and audience
6. EDA routines: 1) frequencies (qualitative) and histograms (quantitative); 2) descriptive/summary statistics; 3) extreme values (outliers); 4) graphing (cross sectional and time series comparisons)
7. Theoretical probability distributions and their applications in statistical modeling
   1) Bernoulli: all zero-one binary variables
   2) Binomial: the sum of independent Bernoulli
   3) Uniform: any variable with equally likely values in a range
   4) Normal (e.g., IQ, height, weight) and lognormal (price, income, and firm size)
   6) Power-law (Pareto): distributions with very large extreme values (the filthy rich, mega city size)

II. Comparison and Correlation (CH4)

1. Relations between two variables: 1) independent; 2) associated/correlated/dependent; 3) causal
2. Correlation coefficient [-1, 1]: 1) positive (>0.3); 2) negative (<-0.3); 3) weak or zero (-0.3, 0.3)
3. Correlation and dependence
   1) Covariance and correlation: measure linear relevance
   2) Independence: joint density = marginal density multiplication
   3) Relation: independence → zero correlation, but not vice versa.
4. Conditional probability P(Y|X) = P(XY)/P(X)=P(Y)P(X|Y)/P(X)
5. Conditional mean E(Y|X) and conditional distribution f(Y|X): depend on X
6. Conditional summary statistics

III. Key Distributions for Hypothesis Testing

   1) Normal and standard normal distributions (applied in the central limit theorem C.L.T.)
   2) t-distribution (testing single coefficient significance; testing equality of two population means)
   3) F-distribution (joint test of regression significance; ratio of two population variance)
   4) Chi-square distribution (test the population variance; normality; maximum likelihood tests)

Lecture 3: Statistical Inference

I. Statistical Inference (CH5)

1. The question for statistical inference: What is the true value of a statistics from the population?
   1) The goal of SI is learning the value of a statistic in the population, e.g., its mean.
   2) Apply statistical inference to uncovering the unknown true value(s) in the population
   3) Define the population of interest, define the population the data represents, and compare the two to assess external validity of our inference
2. Repeated samples: each observed dataset can be viewed as a sample drawn from the population.
3. With repeated samples, the statistics has a distribution and its value differs from sample to sample
4. An estimator is a function or formula for computing the statistics (e.g., sample average).
5. An estimate is the calculated value of the statistics given a particular data sample. Different samples produce different estimates, but they can share the same estimator (e.g., sample average).
6. **Sampling distribution** is the distribution of the estimates of the statistics across many repeated samples of the same size. Standard error SE is the standard deviation of the sampling statistics.
7. **The Law of Large Numbers LLN** (X ~ i.i.d.): a sample average can be brought as close as the average in the population from which it is drawn, by enlarging the sample size. WLLN: Average $(X_i)=(X_1+X_2+\ldots+X_N)/N$ → $E(X_i)$=mu as N goes infinite (sample mean convergence).
   1) Examples: tossing a coin for Prob(head=1); throwing a dice to estimate the mean
   2) Application: how can we estimate a coin's true probability of getting heads if it were biased?
   https://digitalfirst.bfwpub.com/stats_applet/stats_applet_10_prob.html
   https://digitalfirst.bfwpub.com/stats_applet/stats_applet_11_largenums.html
8. **The Central Limit Theorem CLT** (X ~ i.i.d.): in large enough data, the estimated average will be distributed normally around the true expected value, and the variance of this normal distribution is the variance of X over sample size. CLT: **Ave(X)~N(E(X), V(X)/N)**
9. **Key properties of the sampling distribution** (with large sample size N, asymptotic)
   1) Unbiasedness: sample average → population mean as N increases to infinity
   2) Asymptotic normality: the sampling distribution of the sample average is normal
   3) Root-n convergence: the S.E. is inversely proportional to the square root of N
10. Confidence interval (1-alpha) %: How confident are we in the estimated sampling statistics and distribution that represent the true population? The C.I. gives the range of values where we think that the true value falls with a (1-alpha) % likelihood of the sample estimate. Construct the C.I.:
    1) Estimate the parameter (sampling statistics) or the point estimate with the given sample
    2) Calculate standard error using the formula **S.E.(Xbar)=1/sqrt(N)*Std(X)** or bootstrap S.E.
    3) 90% C.I. is 1.6 S.E. interval around the estimate from the data; 95% C.I. 1.96; 99% C.I. 2.6.
    4) Formula: 95% C.I. is [point estimate-1.96*S.E., point estimate+1.96*S.E.]
11. Bootstrap is a resampling method for generating the entire sampling distribution and its statistics.
    1) In practice, we only have one dataset (all the data we have) and we would like to uncover the sampling distribution of samples of the same size to that data.
    2) To estimate standard error of the statistic, we compute the standard deviation of the estimates of the statistics across available bootstrap samples of the same size.
    3) "Pull yourself out of the swamp by your bootstraps." How is that possible?

CH5 Calculation in R: S.E. and C.I. Case Study: What likelihood of loss to expect on a stock portfolio?

Lecture 4: Hypotheses Testing

I. Basic Questions

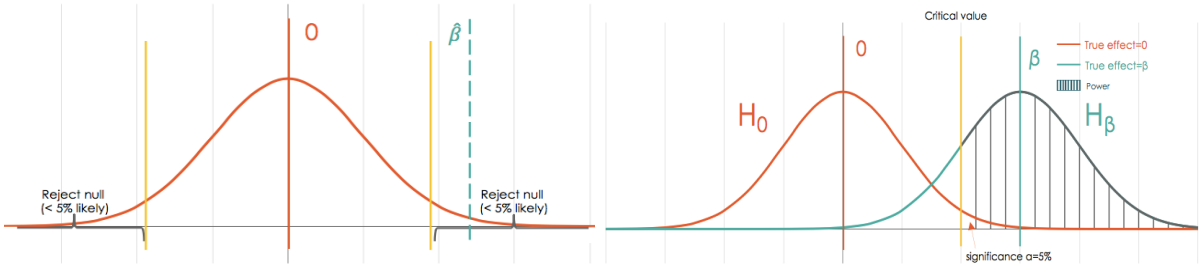
1. Is the sample estimate providing strong enough evidence against/for its population parameter?
2. Examples of HT: 1) is the mean equal to a specific value? 2) is the mean-difference positive?
3. Types of test: 1) one-sided (strictly positive/negative outcomes); 2) two-sided (non-zero outcomes)
4. **A test statistic** is a measure of the distance of the estimated value of the statistics from what its true value would be if the null hypothesis were true. The decision rule in statistical testing is comparing the test statistic to a critical value determined by a chosen significance level.

II. Components of HT

1. null H0 vs alternative H1 (refer only to the population parameter)
2. a test statistic by construction and a benchmark probability distribution
3. significance level given the null is true (the max prob of a false positive being tolerated)
4. critical value in the benchmark distribution (corresponding to a pre-specified significance level)
5. calculate from the data sample for the test statistics and compare it with the chosen critical value

III. Types of errors in HT decisions: 1) FP is type-I error (alpha); 2) FN is type-II error (beta)

HT Decision	Null hypothesis H0	Alternative hypothesis H1
Fail to reject H0	True Negative TN	False Negative FN
Reject H0	False Positive FP	True Positive TP



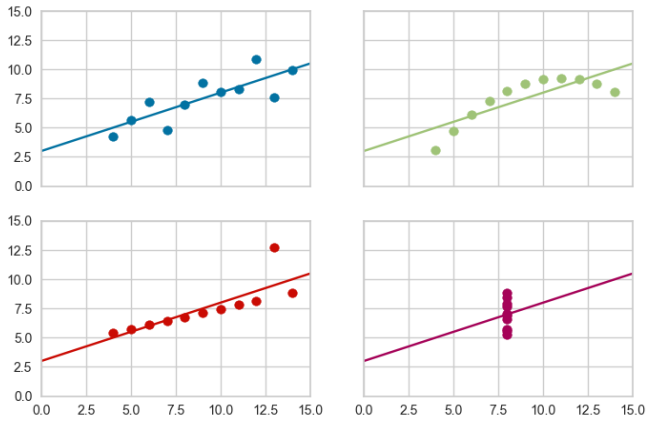
Source: J-PAL. Question: In first-stage rapid testing of COVID-19, which type of error must policy makers avoid?

1. The size of the test is the probability of a false positive (alpha).
2. The level of significance is the max prob of a false positive we can tolerate (pre-chosen). A smaller level of significance (larger critical value) will make it harder to reject the null H0.
3. The power of the test is the probability of avoiding a false negative (1 minus beta).
4. The p-value informs us about the probability of a false positive in sample estimation. The p-value for a test is the smallest significance level at which we can reject the null hypothesis given the value of the test statistics in the sample. Reject the null if the p-value is less than a pre-chosen significance level. Smaller p-value leads to stronger evidence against the null H0.
5. HT (significance test) procedures
   1) Choose the null and alternative hypotheses (mutually exclusive)
   2) Specify the test statistics (t, F, Chi-sq) and its distribution under the null
   3) Select a significance level alpha: Prob(reject $H_0 \mid H_0$) → rejection region
   4) Calculate the sample value of the test statistics (or corresponding p-value)
   5) State the testing conclusion (reject or fail to reject) or refine the test

Lecture 5: Regression Models & Analysis

I. Simple Regression Model (CH7)

1. Function: 1) estimate and quantify the relationship between Y (explained or dependent) and X (explanatory or independent); 2) to forecast (or predict) new Ys out of sample in decision making
2. Definition: Regression analysis is a statistical method that uncovers the average value of Y conditional on X. Regression model is a conditional mean E[Y|X], which is a function of X.
3. Model specification for cross-sectional data: $Y_i=E[Y|X]+e_i$ where e_i is a white noise
4. Simple linear regression model: $Y_i = \alpha+\beta X_i+e_i$ where $e_i \sim$ i.i.d. N(0, sigma^2)
   1) Linearity parameters: a is the intercept coefficient, b is the slope coefficient
   2) Regression line is $Y=\alpha+\beta X=E[Y|X]=E[X]$ because $E[e|X]=E[e]=0$ (e and X are uncorrelated)
   3) Random error/disturbance/shock e_i is normally distributed with E(e)=0 and V(e)=Constant
   4) The ordinary least square (OLS) estimation: minimizing Sum$\{[Yi-E(Y_i|X)]^2\}$
   5) OLS formula for the estimators: b=cov(X, Y)/var(X) and a=E(Y)-b*E(X)
5. Classical OLS assumptions and properties
   1) Linearity, E(e|X) = 0, no multicollinearity, homoskedasticity, no autocorrelation, normality
   2) If OLS assumptions met, then OLS is the Best Linear Unbiased Estimator B.L.U.E.



https://www.scikit-yb.org/en/latest/api/anscombe.html

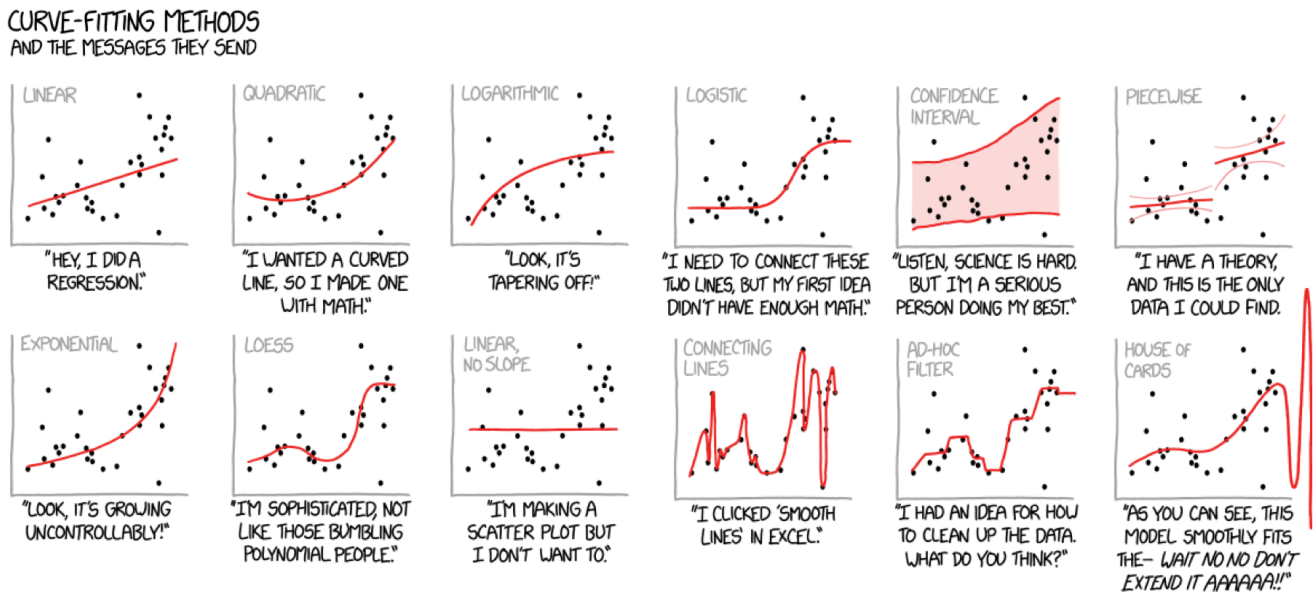
- Among all possibilities, is there a unique line fits exactly all data points?
- How to find a line or curve that can best fit all the data points in a scatter plot?
- What do we mean by best? Is there a standard/measure?
- By minimize the total sum of square errors, the OLS method provides a simple yet "best" solution.

II. Simple Regression Diagnosis (CH7)

1. Measure of the goodness-of-fit: **R-square** (ESS/TSS)
   1) How much variation in Y can be explained by E(Y|X)? Plot E(Y|X) against Y
   2) Variance decomposition: Var(Y)=Var(a+bX)+Var(e), assuming E(e|X)=0
   3) R^2=Var(a+bX)/Var(Y)=1-Var(e)/Var(Y). Range [0, 1]. Adjusted R^2 for dof.
2. Residual test: can we find systematic patterns in the residual?
   1) Residuals Y_i-E($Y_i|X_i$): the difference between Yi and a+bXi
   2) In cross-section, plot e against X: is the conditional variance constant?
   3) In time series, plot e_t against e_{t-1}: is the residuals autocorrelated?
3. Model predictions (interpolation vs extrapolation) and forecast errors (for model comparison)
   1) In-sample estimate E($Y_i|X_i$): given a specific X_i in the sample, what is the predicted Y_i?
   2) Out-of-sample prediction E($Y_i|X_j$): given a new X_i not in the sample, what is the predicted Y_j?
   3) Prediction errors: MSE=sum($[Y_i-E(Y_i|X_i)]^2$)/n, RMSE=sqrt(MSE), MAE=sum$|Y_i-E(Y_i|X_i)|$/n,

IIII. Regression Model Specification (CH8)

1. Functional forms: linear vs nonlinear. Can we transform variables to linear regression?
2. Transformation of variables in levels, in differences, in logs or semi-logs (interpretation vs fit)
3. Influential observations ("outliers"): In EDA, they should be excluded if due to errors
4. Measurement errors (wrong values, recording noise) → biased estimation
5. Why modeling nonlinearity? Points of the story (Helwig, 2021).
   1) Just because a model fits "good" doesn't mean that it is a good model.
   2) Linear regression models may not reveal "true" relationships in the data
   3) More complex models can provide better in-sample fit than the linear model
   4) A "good" model survives out-of-sample test in forecasting (gold standard)
6. Nonlinear regression: polynomial; piecewise linear spline; generalized additive models (GAM)
   1) Polynomial regression captures a certain amount of curvature in a nonlinear relationship.
   2) Splines provide a way to smoothly interpolate between fixed points called knots.
   3) Piecewise linear spline produces connected line segments.
   4) GAM: a technique to automatically fit a spline regression.
7. Nonparametric regression (smoothing): local averaging; local regression; kernel regression.
8. Locally weighted scatterplot smoothing (lowess or loess): a smooth curve fit around a bin scatter.
9. Observation weighted regression (Weighted Least Square – WLS estimation)



https://therbootcamp.github.io/appliedML_2019Jan/_sessions/Fitting/Fitting_practical.html

III. Simple Regression Statistical Inference (CH9)

1. Standard errors of regression coefficient estimates: S.E.(b)=sqrt^{-1}(n)*[S.D.(e)/S.D.(X)]
2. Robust standard error (heteroskedasticity-robust): The White-Huber "sandwich" formula
3. Confidence interval of regression coefficients 95% C.I.(b)=[b-1.96*S.E.(b), b+1.96*S.E.(b)] such that we can expect beta to lie with 95% confidence in the interval estimated by the data
4. Construct and interpret prediction intervals: 95% C.I.(y_p)=[y_p-1.96*S.E.(y_p), y_p+1.96*S.E.(y_p)]
5. Significance test of regression results H0: beta=0. Apply the t-test with p-value reported
6. External validity of a regression analysis: Can we apply the sample result to a new context?
7. Observational data (invalid for causal inference) vs experimental data (controlled effect)

V. Multiple Regression and Statistical Inference (CH10)

1. Why include multiple variables in a linear regression? 1) omitted variables; 2) higher-order terms; 3) interaction between two variables (potentially for higher-order terms)
2. Multiple linear regression model: $Y_i=\alpha+\beta_1X_{1i}+\beta_2X_{2i}+\ldots+b_kX_{ki}+e_i$ where e_i~iid N(0, sigma^2)
3. Significance tests (t) for the coefficient estimates: H0 is alpha=0 and beta=0. The t-statistics.
4. Joint significant test (F) for the regression: H0 is alpha=beta=gamma=…=0. The F-statistics.
5. Confidence interval for the coefficient estimates

VI. Estimation Criteria (the true population parameter is α and its corresponding estimator is a)

1. Unbiasedness: E(a)=α, the expectation of an estimator is equal to its population parameter
2. Consistency: V(a) → 0, as sample size increase to infinity, variance disappears
3. Efficiency: V(a)=<V(a'), smaller variance among competing estimators
4. Note that being unbiased is a precondition for an estimator to be consistent.

Lecture 6: Probabilistic Regressions

How can we estimate and forecast the probability of a certain event Y, Prob(Y)? If we obtain other information related to Y, how can we estimate and forecast the probability of Y happening? (CH11)

I. Linear Probability Model

1. Suppose Y is a binary label to document the occurrence of an event. $Y_0=0$ or $Y_1=1$
2. Unconditional probability $E(Y)=Y_0P(Y=Y_0)+Y_1P(Y=Y_1)=P(Y=1)$
3. Conditional probability $E(Y|X)= Y_0P(Y=Y_0|X)+Y_1P(Y=Y_1|X)=P(Y=1|X)$
4. LPM: $YP_i=P(Y=1|X)=\alpha+\beta_1X_{1i}+\beta_2X_{2i}+\ldots+e_i$ where $e_i \sim$ i.i.d. N(0, sigma^2)
5. Disadvantage: The fitted values/line of YP_i can go out of probability bound [0, 1]
6. Solution: change the linearity to a probability function, as in the logit and probit

II. Logit and Probit Models

1. The Logit: $YP_i=P(Y=1|X)=G(X)=\exp(\alpha+\beta_1X_{1i}+\beta_2X_{2i}+\ldots)/1+\exp(\alpha+\beta_1X_{1i}+\beta_2X_{2i}+\ldots)$
2. The Probit: $YP_i=P(Y=1|X)=F(X)=Prob(Z=<\alpha+\beta_1X_{1i}+\beta_2X_{2i}+\ldots)$ where $Z \sim$ iid N(0, 1)
3. Both G(.) and F(.) are cumulative probability distribution functions (nondecreasing up to one)
4. Logistic regression does not assume the residuals are normally distributed nor constant variance
5. Estimation method: maximum likelihood numerical optimization
6. The predicted $YP= Prob(Z=a+b_1X_{1i}+b_2X_{2i}+\ldots+b_kX_{ki}<z)$

III. Statistical Inference and Model Evaluation

1. Peuso-R^2 (McFadden) goodness of fit: $1-\ln LM_1/\ln LM_0$ where LM_0 is the intercept-only model
2. Residual assessment: the deviance residual is useful for checking if individuals are not well fit
3. Likelihood ratio (LR) test: whether the observed difference in model fit is statistically significant
4. Prediction errors, classification errors, confusion matrix, receiver operating characteristics (ROC)

Reference: Logistic Regression https://uc-r.github.io/logistic_regression

Lecture 7: Time Series Regression

Time series analysis comprises methods for analyzing time series data in order to extract meaningful statistics and other characteristics of the data. Time series forecasting is the use of a model to predict future values based on previously observed values (Wikipedia).

I. Time Series Data and Regressions

1. Data frequency: intra-day, daily, weekly, monthly, quarterly, semi-annually, annually
2. Regression model: $Y_t=\alpha+\beta X_t+e_t$ or $\ln Y_t= \alpha+\beta \ln X_t+e_t$ or $\Delta Y_t= \alpha+\beta \Delta X_t+e_t$
3. Autoregression model: $Y_t=\alpha+\beta_1 Y_{t-1}+\beta_2 Yt_{-2}+\ldots+e_t$ where e_t is a white noise
4. Auto-Regressive Integrated Moving Average: ARIMA(p, d, q)

II. Trends, Seasonality, Structural Breaks, Random Walks

1. Common trend between X and Y → spurious correlation/regression
2. Seasonality in levels or changes → spurious correlation/regression
3. Random-walk variables → biased estimates (S.E. and C.I.) and unreliable inference
4. Detection: 1) time series plot; 2) structural breaks test; 3) unit root tests for random walks
5. Solutions: 1) model trend as a function of time; 2) model seasonality as dummy variables; 3) remove random walks by first or second differencing 4) separate the sample by the breaks

III. Serial Correlation (Autocorrelation)

1. Problem: If Y_t is serially correlated, the usual standard errors of regressions are wrong
2. Detection: 1) plot Y_t against time; 2) plot Y_t against Y_{t-1}; 3) formal statistical tests
3. Solutions: 1) Newey-West S.E. (robust); 2) Lagged dependent variable $Y_t= \alpha+\beta X_t+\gamma Y_{t-1}+e_i$

IV. Procedures in Time Series Modeling

1. Plot the series, visualize the patterns, and perform statistical tests
2. Handle trends by transforming variables in levels, changes, relative changes
3. Model trend (include a function of time) and seasonality (include dummies)
4. Handle serial correlation: include or don't include lags of the explanatory variables
5. Interpret coefficients in a way that pays attention to potential trend and seasonality

Reference

Forecasting: Principles and Practice
https://otexts.com/fpp2/
https://otexts.com/fpp2/arima.html
https://otexts.com/fpp2/graphics.html
https://otexts.com/fpp2/regression.html
https://otexts.com/fpp2/decomposition.html
https://rpubs.com/riazakhan94/arima_with_example
https://search.r-project.org/CRAN/refmans/EnvStats/html/serialCorrelationTest.html

Appendix: Case Study

CH1
1. How to find a good deal among hotels? Data collection: web scraping
2. How much online and offline prices differ?
3. How to measure the quality of corporate management?

CH2
1. Does immunization save lives?
2. How to identify successful football managers in the football league?

CH04A Management quality and firm size: describing patterns of association
Are larger companies better managed? What is the relationship between firm size and the quality of management?

CH5 Case Study: What likelihood of loss to expect on a stock portfolio?

CH7 Case Study: Where are underpriced hotels (for their location and quality)?

CH11 Case Study: Does smoking pose a health risk?

CH01A Finding a good deal among hotels: data collection
CH01B Comparing online and offline prices: data collection
CH01C Management quality: data collection

CH02A Finding a good deal among hotels: data preparation
CH02B Displaying immunization rates across countries
CH02C Identifying successful football managers

CH03A Finding a good deal among hotels: data exploration
CH03B Comparing hotel prices in Europe: Vienna vs London
CH03C Measuring home team advantage in football
CH03D Distributions of body height and income

CH04A Management quality and firm size: describing patterns of association

CH05A What likelihood of loss to expect on a stock portfolio?

CH06A Comparing online and offline prices: testing the difference
CH06B Testing the likelihood of loss on a stock portfolio

CH07A Finding a good deal among hotels with simple regression

CH08A Finding a good deal among hotels with non-linear function
CH08B How is life expectancy related to the average income of a country?
CH08C Measurement error in hotel ratings

CH09A Estimating gender and age differences in earnings
CH09B How stable is the hotel price–distance to center relationship?

CH10A Understanding the gender difference in earnings
CH10B Finding a good deal among hotels with multiple regression

CH11A Does smoking pose a health risk?
CH11B Are Australian weather forecasts well-calibrated?

CH12A Returns on a company stock and market returns
CH12B Electricity consumption and temperature

CH13A Predicting used car value with linear regressions

CH14A Predicting used car value: log prices
CH14B Predicting AirBnB apartment prices: selecting a regression model

CH15A Predicting used car value with regression trees

CH16A Predicting apartment prices with random forest

CH17A Predicting firm exit: probability and classification

CH18A Forecasting daily ticket sales for a swimming pool
CH18B Forecasting a house price index

CH19A Food and health

CH20A Working from home and employee performance
CH20B Fine tuning social media advertising

CH21A Founder/family ownership and quality of management

CH22A How does a merger between airlines affect prices?

CH23A Import demand and industrial production
CH23B Immunization against measles and saving children

CH24 Estimating the effect of the 2010 Haiti earthquake on GDP
CH24 Estimating the impact of replacing football team managers