Syllabus of the course Data Analytics for the Swedish Program

In this course, students will learn the fundamentals of Econometrics, with a focus on cross-sectional data, and receive an introduction to Machine Learning (ML). The course also covers essential mathematical methods used in economic and financial analysis, emphasizing the role of data science in business and economic decision-making.

Key econometric concepts include random sampling, the central limit theorem, estimation, statistical inference (hypothesis testing), regression models, causality, and randomized experiments. ML topics include supervised learning (prediction and classification) and unsupervised learning (clustering).

The R programming language will be used throughout the course to support learning in both Econometrics and ML. Students will work with various R packages to analyze regression models, assess causality, design and evaluate randomized experiments, and perform prediction, classification, and clustering tasks. They will also learn effective techniques for visualizing data and communicating results.

Intended Learning Outcomes

By the end of the course, students will be able to:

- Apply sampling, estimation, and hypothesis testing techniques to economic and business data.
- Build, interpret, and evaluate simple and multiple regression models.
- Apply key machine learning methods including supervised (regression, classification) and unsupervised learning (clustering, dimensionality reduction).
- Use R and relevant packages for data visualization and modeling.
- Critically evaluate models and interpret results in context.
- Communicate analytical results effectively.

Material

A comprehensive slide package (PDF) will be provided, summarizing the chapters covered in the course. Most of the R code required to complete the assigned exercises will also be provided.

Readings

Readings in Statistics for Business and Economics (SBE).

- Lecture 1. *Sampling and Sampling Distributions*. Readings Ch. 6 in SBE (pp. 248-287). Sections 6.1-6.4
- Lecture 2. *Point Estimation*. Readings Ch. 7 in SBE (pp. 288-331). Sections 7.1-7.2
- Lecture 3. *Point Estimation*. Readings Ch. 7 in SBE (pp. 288-331). Sections 7.3-7.7
- Lecture 4. *Estimation: Additional Topics*. Readings Ch.8 in SBE (pp. 332-349). Sections 8.1-8.3
- Lecture 5. *Hypothesis Testing: Singe Population*. Readings Ch.9 in SBE (pp. 350-388). Section 9.1
- Lecture 6. *Hypothesis Testing for the Mean*. Readings Ch.9 in SBE (pp. 350-388). Sections 9.2, 9.5, 9.6

- Lecture 7. *Hypothesis Testing for the Mean*. Readings Ch.9 in SBE (pp. 350-388). Sections 9.3, 9,4, 9.6
- Lecture 8. *Hypothesis Testing: Additional Topics*. Readings Ch.10 in SBE (pp. 389-420). Sections 10.1 – 10.5
- Lecture 9. *Linear Regression Model*. Readings Ch.11 in SBE (pp. 421-476). Sections: 11.1,11.7
- Lecture 10. *Linear Regression Model*. Readings Ch.11 in SBE (pp. 421-476). Sections: 11.2-11.6
- Lecture 11. *Multiple Regression Model*. Readings Ch.12 in SBE (pp. 477-554). Sections: 12.1 - 12.8.
- Lecture 12. *Machine Learning*. Sections in "Introduction to Data Science": 27.1, 27.4.1- 27.4.6, 27.4.8, 31.1 (Regression), 31.3 (Logistic Regression), and 31.5 (K-Nearest Neighbors)

Literature

- "Statistics for Business and Economics," 10th Global edition 2022, by Paul Newbold, William L. Carlson and Bett M. Thorne.

- "Introduction to Data Science - Data Wrangling and Visualization with R," 2023, by Rafael A. Irizarry (http://rafalab.dfci.harvard.edu/dsbook-part-1/)

- "Advanced Data Science - Statistics and Prediction Algorithms Through Case Studies," 2023, by Rafael A. Irizarry (http://rafalab.dfci.harvard.edu/dsbook-part-2/)

Examination

Written exam (max 65p) and project (max 35p); standard grades apply: A+, A, A-, B+, B, B-, C+, C, C-, D+, D, D-, and F (fail).

Date and deadline for the exam and project: **TBA**.

Course Project

As part of the course, students will complete an applied data analytics project using the R programming language. The project is divided into three integrated parts:

1. **Marketing Case**. Analysis of survey data to explore customer characteristics, compute confidence intervals, and compare group means, providing actionable insights for targeted marketing strategies.

2. **Global Challenge Case.** Multiple linear regression modeling of life expectancy data from diverse countries to identify key socio-economic, health, and environmental predictors, linking analytics to global development issues.

3. **Credit Risk Case**. Application of machine learning classification techniques (K-Nearest Neighbors and Logistic Regression) to predict credit default risk, evaluating model performance with accuracy, sensitivity, specificity, and F1-scores.

This project emphasizes real-world data handling, statistical inference, regression analysis, and supervised learning methods. Students will be assessed on the accuracy of their analyses and the clarity of their interpretations.

Typical Problem Types to be Mastered

Lecture 2

- Show that the sample mean, proportion, and variance all are unbiased estimators of their population counterparts.
- Examples of biased estimators.
- Sample size determination polls.
- Calculating confidence intervals for expected returns and risk for fund A.

Lecture 3

- CI for the difference in expected returns fund A and fund B.
- CI for the ratio of volatilities fund A and fund B.
- CI for the difference in proportions US election 2017 Donald Trump and Hillary Clinton.
- CI for examining treatment effects (e.g., examining effects of tax cuts).

Lecture 4

- Test the hypothesis that IQ score is more than 100.
- Test the hypothesis that expected returns of fund A is more than 5%.
- Test the hypothesis that the risk of fund A is more than 1%.
- Calculate the power of the IQ score test when the true IQ score is 105.
- Calculate the *p*-value of the IQ score test.

Lecture 5 - 8

- Test the hypothesis that expected returns of fund A are equal to the expected returns of fund B.
- Test the hypothesis that fund B is four times as risky as fund A.
- Test the hypothesis that tax cuts are not having any impact on household consumption.
- Test the hypothesis that the proportion of Hillary Clinton supporters are equal to the proportion of Donald Trump supporters.
- Calculate the power of all tests above.

Lecture 9 and 10

L9.1: Test the hypothesis that Performance and Aptitude are positively associated. Use the Aptitude and Performance data given below.

L10.1: In a SLR model Performance is selected as the dependent variable (Y_i) and Aptitude is selected as the independent variable (x_i). The same data as is in L9.1 should be used for this exercise.

i. Estimate this model (i.e., calculate b_0, b_1 and s_e^2).
ii. Interpret the estimated marginal effect.
iii. Calculate a 90% confidence interval for β_1. Interpret this interval.
iv. Test the hypothesis that Aptitude has a positive effect on Performance. Use a 5% significance level. What can you say about the *p*-value of this test?
v. Calculate SST, SSR, SSE, R^2-value, and the adjusted R^2-value. Comment on the R^2-value.
vi. Calculate the predicted performance level for a new candidate observation x = 72. Also calculate a 95% prediction interval for the performance level of this candidate.

L10.2: Verify the answers above by running the R-code for SLR and MLR models.

Lecture 11

L11.1: In an MLR model Performance is selected as the dependent variable (Y_i) and Aptitude is selected as a first independent variable (x_{1i}) and Personality is selected as a second independent variable (x_{2i}). The data for these variables are given below.

i. Using the estimates of b_1 and b_2 in L11 (p.3), interpret the estimated marginal effects.
ii. Calculate an 90% confidence interval for β_1 and β_2. Interpret these intervals.
iii. Test the hypothesis that Aptitude has a positive effect on Performance. Test the hypothesis that Personality has a positive effect on Performance. Lastly, test the overall significance of this model. Use a 5% significance level for all tests. What can you say about the *p*-values of these tests?
iv. Calculate SST, SSR, SSE, R^2-value, and the adjusted R^2-value. Comment on the adjusted R^2-value ($\underline{R_2}$).
v. Calculate the predicted performance level for a new candidate with observations x_1 = 72 and x_2 = 14. Compare this prediction to the previous one in L10.1.

L11.2: Verify the answers above by running the R-code for SLR and MLR models.

L11.3: In an MLR model Wage is selected as the dependent variable (Y_i) and Education (x_{1i}), Experience (x_{2i}), and Tenure (x_{3i}) are selected as independent variables. All data and estimation results needed to answer the questions in this exercise are obtained by running the R-code for SLR and MLR models.

i. Estimating this model, interpret the estimated marginal effects b_1, b_2, and b_3. Which variable has the strongest impact on Wage?
ii. Test the hypothesis that the variables are individually significant. Also, test the overall significance of this model. Use a 5% significance level for all tests.
iii. Calculate SST, SSR, SSE, R^2-value, and the adjusted R^2-value. Comment on the adjusted R^2-value.

L11.4: Examine if the square of the independent variables should be added as a sub-set of regressors to the model in L12.1. You may use the notation: $z_{1i} = x_{1i}^2$ (the square of Education), $z_{2i} = x_{2i}^2$ (the square of Experience), and $z_{3i} = x_{3i}^2$ (the square of Tenure).

L11.5: Test the hypothesis that there is a systematic difference in wages between females and males for the same level of experience, by adding the dummy variable Female (x_{4i}) to the SLR model with Wage (Y_i) as dependent variable and Experience (x_{2i}) as independent variable. Also, test the hypothesis that the wage trend with respect to experience depends on the gender. Use a 5% significance level for both tests.

L11.6: Nonlinear regressions. In an MLR model Performance is selected as the dependent variable (Y_i) and Aptitude is selected as a first explanatory variable (x_{1i}) and the square of Aptitude is selected as a second explanatory variable (x_{1i}^2). For this exercise, the Aptitude2 and the Performance2 data below are used.

i. Using the estimates of b_1 and b_2 in L12 (p.4), interpret the estimated marginal effects for min, mean, median, and max values of Aptitude.
ii. For which value of Aptitude is the estimated marginal effect on Performance equal to zero (this yields the value of Aptitude that maximizes Performance)?

L11.7: Transformations for nonlinear regression models.

i. Estimate the Cobb-Douglas production function for US data (see L12 pp.7-12). Interpret the estimates. State the estimated model. For this exercise, use the US production data file "USProdFuncData.dta" that I sent to you by mail.
ii. Test the constant return to scales hypothesis: $\beta_1 + \beta_2 = 1$.

Data for Exercises Lecture 2 - 8

- Returns fund Data A (%): 6.7, 6.5, 6.8, 5.2, 5.0, 5.1, 5.5, 5.5, 5.6, 5.9, 6.1, 6.0.
  Returns fund Data B (%): 15.5, 14.9, 15.1, 5.6, 1.1, 4.2, 5.7, 6.1, 11.4, 12.4, 11.9, 10.5.

- Poll data US election 2017: Hillary Clinton 467 positive votes of 1000. Donald Trump 448 positive votes of 1000.

- Household consumption data (SEK) before and after tax cuts:

  X_i (before): 4567 3891 5823 10915 3475 7763 4511 5006 3299 4654
  Y_i (after): 4587 3999 6002 10956 3344 7780 4599 5111 3031 4670

- IQ score data: 98, 85, 111, 112, 101, 137, 81, 95, 100, 105, 121, 88, 99, 102, 110, 107.

Some Data for Exercises Lecture 9 - 11

- Aptitude: 45, 81, 65, 87, 68, 91, 77, 61, 55, 66, 82, 93, 76, 83, 61, 74
  Personality: 9, 15, 11, 15, 14, 19, 12, 10, 9, 14, 15, 14, 16, 18, 15, 12
  Performance: 56, 74, 56, 81, 75, 84, 68, 52, 57, 82, 73, 90, 67, 79, 70, 66

- Aptitude2: 52, 56, 56, 57, 66, 67, 68, 70, 73, 74, 75, 79, 81, 82, 84, 90
  Performance2: 50, 61, 63, 64, 72, 78, 69, 77, 75, 71, 74, 64, 62, 61, 59, 50