in collaboration with

Faculty of Engineering, Environment and Computing
7089 CEM: Introduction to Statistical Methods for Data Science

Assignment Brief

Module Title Introduction to Statistical Methods for Data Science	Individual	Cohort (Sept/Jan): Sept and Jan start	Module Code 7089 CEM
Coursework Title **Modeling EEG signals using polynomial regression**			Hand out date: TBA
Lecturer Uttam Pokharel			Due date and time: TBA
Estimated Time (hrs.): 4 weeks Word Limit*: 2000- 2500	Coursework type: Individual assignment		% of Module Mark: 100%

- Submission arrangement: Online via SchoolWorkshop pro: https://schoolworkspro.com.
- File types and method of recording: Report (Word), Programme code (R, or Python/Matlab script)
- Mark and Feedback date: 2 weeks after submission
- Mark and Feedback method (e.g. in lecture): provided in SchoolWorkshop pro.

Module Learning Outcomes Assessed:

- Demonstrate knowledge of underlying concepts in probability and statistics used in Data Science.
- Select and apply appropriate statistical methods or techniques to solve problems or analyze data sets.
- Use modern software to solve real world problems and analyze large data sets.
- Interpret the results of their analyses and communicate those results accurately.

Task and Mark distribution:

Coursework Description:

The aim of this assignment is to select the best regression model (from a candidate set of nonlinear regression models) that can well describe the relationship between several 'simulated' brain electroencephalogram (EEG) signals. EEG is a widely used non-invasive method to record electrical activity of the brain. Compared with other neuroimage data, such as functional magnetic resonance imaging (fMRI), EEG technique is much cheaper and has excellent temporal resolution. As a result, EEG-based analysis and modeling approaches have been extensively applied to characterize various neurological disorders (e.g. Parkinson's disease, epilepsy, tremor) and the development of brain-computer interface. To achieve those goals, it is very important to understand the connectivity between different brain areas, which can be obtained through the modeling and analysis of different EEG channels.

Data sets : Provided in https://schoolworkspro.com.

The 'simulated' 5 EEG time-series data are provided in the two separate excel files. The first X.csv file contains 4 input EEG signals **X** = {x_1, x_2, x_3, x_4}, respectively; and the second y.csv file contains the output EEG signal **y**. The file time.csv contains the sampling time of all EEG data in seconds. All the 5 EEG signals are subject to additive noise (assuming independent and identically distributed ("i.e.,") Gaussian with zero-mean) with unknown variance.

Task 1: Preliminary data analysis

You should first perform an initial exploratory data analysis, by investigating:

- Time series plots (of input and output EEG signals)
- Distribution for each EEG signal
- Correlation and scatter plots (between different input EEG signals and the output EEG) to examine their dependencies

Task 2: Regression – modeling the relationship between EEG signals

We would like to determine a suitable mathematical model in explaining the relationship between the input EEG signals and the output signal, assuming such a relationship (brain connectivity) can be described by a polynomial regression model. Below are 5 candidate nonlinear polynomial regression models, and only one of them can 'truly' describe such a relationship. The objective is to identify this 'true' model from those candidate models following Tasks 2.1 – 2.6.

Candidate models are with the following structures:

Model 1: $y = \theta_1\, x_4 + \theta_2\, {x_1}^2 + \theta_3\, {x_1}^3 + \theta_4\, {x_2}^4 + \theta_5\, {x_1}^4 + \theta_{bias} + \varepsilon$

Model 2: $y = \theta_1\, x_4 + \theta_2\, {x_1}^3 + \theta_3\, {x_3}^4 + \theta_{bias} + \varepsilon$

Model 3: $y = \theta_1\, {x_3}^3 + \theta_2\, {x_3}^4 + \theta_{bias} + \varepsilon$

Model 4: $y = \theta_1\, x_2 + \theta_2\, {x_1}^3 + \theta_4\, {x_3}^4 + \theta_{bias} + \varepsilon$

Model 5: $y = \theta_1\, x_4 + \theta_2\, {x_1}^2 + \theta_3\, {x_1}^3 + \theta_4\, {x_3}^4 + \theta_{bias} + \varepsilon$

Task 2.1:
Estimate model parameters $\boldsymbol{\theta} = \{\theta_1, \theta_2, \cdots, \theta_{bias}\}^T$ for every candidate model using Least Squares ($\widehat{\boldsymbol{\theta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$), using the provided input and output EEG datasets (use all the data for training).

Task 2.2:
Based on the estimated model parameters, compute the **model residual (error) sum of squared errors (RSS)**, for every candidate model.

$$RSS = \sum_{i=1}^{n} \left(y_i - \mathbf{x}_i\widehat{\boldsymbol{\theta}}\right)^2$$

Here $\mathbf{x}_i$ denotes the i^{th} row (i^{th} data sample) in the input data matrix $\mathbf{X}$, $\widehat{\boldsymbol{\theta}}$ is a column vector.

Task 2.3:
Compute the **log-likelihood function** for every candidate model:

$$\ln p\left(D|\widehat{\boldsymbol{\theta}}\right) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\hat{\sigma}^2) - \frac{1}{2\hat{\sigma}^2}\mathrm{RSS}$$

Here, $\hat{\sigma}^2$ is the variance of a model's residuals (prediction errors) distributions $\hat{\sigma}^2 = \mathrm{RSS}/(n-1)$, with n the number of data samples.

Task 2.4:
Compute the **Akaike information criterion (AIC)** and **Bayesian information criterion (BIC)** for every candidate model:

$$AIC = 2k - 2\ln p\left(D|\widehat{\boldsymbol{\theta}}\right)$$
$$BIC = k \cdot \ln(n) - 2\ln p\left(D|\widehat{\boldsymbol{\theta}}\right)$$

Here $\ln p\left(D|\widehat{\boldsymbol{\theta}}\right)$ is the log-likelihood function obtained from Task 2.3 for each model, k is the number of estimated parameters in each candidate model.

Task 2.5:
Check the distribution of model prediction errors (residuals) for each candidate model. Plot the error distributions, and evaluate if those distributions are close to Normal/Gaussian (as the output EEG has additive Gaussian noise), e.g., by using Q-Q plot.

Task 2.6:
Select ‘best’ regression model according to the AIC, BIC and distribution of model residuals from the 5 candidate models, and explain why you would like to choose this specific model.

Task 2.7:
Split the input and output EEG dataset (**X** and **y**) into two parts: one part used to train the model, the other used for testing (e.g. 70% for training, 30% for testing). For the selected ‘best’ model, 1) estimate model parameters use the training dataset; 2) compute the model’s output/prediction on the testing data; and 3) also compute the 95% (model prediction) confidence intervals and plot them (with error bars) together with the model prediction, as well as the testing data samples.

Task 3: Approximate Bayesian Computation (ABC)
Using ‘rejection ABC’ method to compute the posterior distributions of the ‘selected’ regression model parameters in Task 2.

1) You only need to compute 2 parameter posterior distributions -- *the 2 parameters with largest absolute value in* your least squares estimation (Task 2.1) of the selected model. Fix all the other parameters in your model as constant, by using the estimated values from Task 2.1.

2) Use a Uniform distribution as prior, around the estimated parameter values for those 2 parameters (from the Task 2.1). You will need to determine the range of the prior distribution. 3) Draw samples from the above Uniform prior, and perform rejection ABC for those 2 parameters. 4) Plot the joint and marginal posterior distribution for those 2 parameters.

5) Explain your results.

Marking Scheme
This coursework is worth 15 credits (100%). This will be marked according to:
• 20% will be given for performing
Task 1: Preliminary data analysis (histogram plots, simple input output correlation measures, time series plots, fitting linear model, …). If you create any programming code, you must include this in the report.
• 40% will be given for performing
Task 2: Regression: Task 2.1 – Task 2.6, 5% each; Task 2.7 has 10%. • 20% will be given to perform the rejection Approximate Bayesian computation (ABC) to compute the (approximated) posterior distribution of the regression model parameters.
• 10% will be given to appropriate discussion and interpretation of the results you obtained.

• 10% will be awarded for writing the report (around 3000-4000 words) in a structured, readable form and submitting the executable R scripts. The report should be in sections with appropriate headings, and should include introduction, results, discussion and conclusion sections.

• All your programming code should be included in the Appendix of your report. Please display them in a structured way (put headings like Task 1, Task 2.3, etc.), with appropriate comments/annotations. You need to attach the original R code (or Python/Matlab), **NOT** the screenshots of the code. The code will be marked as part of the above marking scheme (for all the Tasks in this coursework, you will need to provide the corresponding code; when you describe/discuss the Tasks in the main text of the report, please reference the corresponding code section in the Appendix).

Notes:

1. You are expected to use the Coventry University Harvard Referencing Style. For support and advice on this student can contact Centre for Academic Writing (CAW).
2. Please notify your registry course support team and module leader for disability support.
3. Any student requiring an extension or deferral should follow the university process as outlined here.
4. The University cannot take responsibility for any coursework lost or corrupted on disks, laptops or personal computer. Students should therefore regularly back-up any work and are advised to save it on the University system.
5. If there are technical or performance issues that prevent students submitting coursework through the online coursework submission system on the day of a coursework deadline, an appropriate extension to the coursework submission deadline will be agreed. This extension will normally be 24 hours or the next working day if the deadline falls on a Friday or over the weekend period. This will be communicated via your Module Leader.
6. Assignments that are more than 10% over the word limit will result in a deduction of 10% of the mark i.e., a mark of 60% will lead to a reduction of 6% to 54%. The word limit includes quotations, but excludes the bibliography, reference list and tables.
7. You are encouraged to check the originality of your work by using the draft Turnitin links on your Moodle Web.
8. Collusion between students (where sections of your work are similar to the work submitted by other students in this or previous module cohorts) is taken extremely seriously and will be reported to the academic conduct panel. This applies to both coursework and exam answers.
9. A marked difference between your writing style, knowledge and skill level demonstrated in class discussion, any test conditions and that demonstrated in a coursework assignment may result in you having to undertake a Viva Voce in order to prove the coursework assignment is entirely your own work.

10. If you make use of the services of a proof reader in your work you must keep your original version and make it available as a demonstration of your written efforts.

11. You must not submit work for assessment that you have already submitted (partially or in full), either for your current course or for another qualification of this university, unless this is specifically provided for in your assignment brief or specific course or module information. Where earlier work by you is citable, i.e. it has already been published/submitted, you must reference it clearly. Identical pieces of work submitted concurrently will also be considered to be self-plagiarism.

Mark allocation guidelines to students:

0-39	40-49	50-59	60-69	70+	80+
Work mainly incomplete and /or weaknesses in most areas	Most elements completed; weaknesses outweigh strengths	Most elements are strong, minor weaknesses	Strengths in all elements	Most work exceeds the standard expected	All work substantially exceeds the standard expected

Marking Rubric

Task 1 – Task 3 (80%)

< 40%	40-49%	50-59%	60-69%	70+%
Little or no implementation of the Tasks using R (or other programming languages) and required approach. Did not describe all the steps in a clear and structured way. Programming code is only partially or not included in the Appendix. It is not displayed in a structured way with explicit annotations. The code is not referenced appropriately in the main text. Some or little results are presented quantitatively. A lack of use of figures and tables.	Some implementation of the Tasks using R (or other programming languages), with or without use of required approach. Partially described the steps of the implementation. Some programming code is included in the Appendix. It is not displayed in a structured way or without explicit annotations. The code is not referenced appropriately in the main text. Some results are presented quantitatively, with or without the use of figures and tables.	Good implementation of the Tasks using required approach and R (or other programming languages). Describe all the steps in a clear and structured way. All programming code is included in the Appendix, displayed in a structured way with clear annotations, and is referenced appropriately in the main text. Results are presented quantitatively and clearly, with the use of figures and tables.	Very good implementation of the Tasks using required approach and R (or other programming languages). Describe all the steps in a clear and structured way. All programming code is included in the Appendix, displayed in a structured way with very clear annotations, and is referenced appropriately in the main text. Results are well presented quantitatively and clearly, with the use of figures and tables.	Excellent implementation of the Tasks using exactly the required approach and R (or other programming languages). Describe all the steps in a clear and structured way. All programming code is included in the Appendix, displayed in a structured way with excellent annotations, and is referenced accurately in the main text. Results are excellently presented and evaluated quantitatively, with the use of figures and tables.
Discussion and Interpretation (10%)				
Little or no interpretation of the results; without appropriate discussions and reflections.	Some interpretation of the results, but little in-depth discussions and reflections.	Good interpretation of the results, with appropriate discussions and reflections.	Very good interpretation of the results, with extensive discussions and reflections.	Excellent interpretation of the results, with in-depth discussions and reflections.
Report writing (10%)				
The report is poorly written without a structured, readable format. A lack of clear presentation and interpretation	The report is written in a readable format but without a clear structure. A lack of clear presentation and interpretation of figures and tables.	The report is written in a structured, readable format, with clear display and interpretation of figures/tables.	The report is well written in a structured, readable format, with clear display and interpretation of figures/tables.	The report has an excellent presentation. It is written in a structured, readable format, with apparent display and interpretation of figures/tables.

of figures and tables.				