Softwarica COLLEGE OF IT & E-COMMERCE | in collaboration with Coventry University

Student Assignment Brief

Contents

Assignment Information

Module Name:	Introduction to Statistical Methods for Data Science
Module Code:	7089 CEM
Assignment Title:	Modeling Blood Glucose Dynamics Using Nonlinear Regression
Assignment Due:	Dec 12 23:59PM 2025

Assignment Credit:	20 credits
Word Count:	3000-4000 words
Assignment Type:	Coursework
Grading:	Percentage Grade

Assessment Overview

You will be provided with an overall grade between 0% and 100%. You have one opportunity to pass the assignment at or above 40%.

Important Notice

The work you submit for this assignment must be your **own independent work**. More information is available in the 'Assignment Task' section of this assignment brief.

Assessed Module Learning Outcomes

The Learning Outcomes for this module align with the marking criteria which can be found at the end of this brief. Ensure you understand the marking criteria to ensure achievement of the assessment task. On completion of this module, the student should be able to:

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

Assignment Task

Coursework Description:

In this assignment, your objective is to build and compare nonlinear regression models that can predict the future blood glucose level (bg+1:00) of a Type 1 Diabetes (T1D) patient using a subset of numerical features derived from the patient's recent time-series physiological and lifestyle data. This analysis utilizes the BRIST1D dataset, available on Kaggle.

The goal is to identify the most suitable model among five nonlinear polynomial regression candidates using statistical criteria such as RSS, AIC, BIC, log-likelihood, and residual diagnostics.

Dataset Description

The dataset contains time-series data collected from T1D patients, comprising sensor-based, manually recorded, and smartwatch-based signals. We focus only on numerical predictors appropriate for regression analysis.

Target Variable:

bg+1:00 (Y): Blood glucose reading in mmol/L one hour into the future

Selected Predictor Variables (Numerical and Time-Series Only):

bg_mean (x1): Average of all available past glucose readings (e.g., bg-0:05 to bg-5:45), representing recent historical blood glucose levels.

insulin_sum(x2): Total insulin dose received over the historical window, calculated by summing all insulin-XX columns.

carbs_sum(x3): Total carbohydrate intake over the time window, obtained by summing all carbs-XX columns.

hr_mean(x4): Average heart rate over the past few hours, calculated using hr-XX columns.

steps_sum(x5): Total steps taken during the period, summed from steps-XX columns.

cals_sum(x6): Total calories burned over the time window, computed by summing cals-XX columns.

Categorical variables like activity-X: XX and unprocessed time values are excluded.

Experimental Setup & Data Format

To facilitate comparability with previous studies, the dataset has been shuffled five times to allow 5x2-fold cross-validation (CV). Each shuffle undergoes 2-fold CV, generating 10 performance measurements for statistical evaluation.

The dataset is available in three different file formats:

dataset.csv

Task 1: Preliminary data analysis:

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

• Time series plots (of input and output signal)

• Distribution for each variable

• Correlation and scatter plots (between different combinations of input and output variables) to examine their dependencies

Task 2: Regression – modelling the relationship between gene expressions

We aim to predict `y = bg+1:00` using the selected features. *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.*

To accomplish these objectives, understanding the interconnection between different variables is crucial, which can be achieved through modeling and analyzing the provided data.

Data sets: Provided in https:/c4mpus.com.

Candidate models are with the following structures:

**Model 1: **
$y = \beta_1 \cdot x_1^3 + \beta_2 \cdot x_2^2 + \beta_3 \cdot x_3^2 + \beta_4 \cdot x_4 + \beta_5 \cdot x_5 + \beta_6 \cdot x_6 + \beta_0$
Model 2:
$y = \beta_1 \cdot x_1^2 + \beta_2 \cdot x_2^2 + \beta_3 \cdot x_3^3 + \beta_4 \cdot x_4 + \beta_5 \cdot x_5 + \beta_6 \cdot x_6 + \beta_0$
Model 3:
$y = \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \beta_3 \cdot x_3 + \beta_4 \cdot x_4^2 + \beta_5 \cdot x_5 + \beta_6 \cdot x_6^2 + \beta_0$
Model 4:
$y = \beta_1 \cdot x_1^2 + \beta_2 \cdot x_2^2 + \beta_3 \cdot x_3^2 + \beta_4 \cdot x_4^2 + \beta_5 \cdot x_5^2 + \beta_6 \cdot x_6^2 + \beta_0$
Model 5:
$y = \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \beta_3 \cdot x_3 + \beta_4 \cdot x_4 + \beta_5 \cdot x_5 + \beta_6 \cdot x_6 + \beta_7 \cdot x_1 x_2 + \beta_8 \cdot x_3 x_4 + \beta_9 \cdot x_2 x_6 + \beta_0$

Task 2.1:

Estimate model parameters $\theta = \{\theta_1, \theta_2, \dots \dots, \theta_{bias}\}^T$ ***for every candidate model using Least Squares:*** $(\hat{\theta} = (X^T X)^{-1} X^T y)$ ***using the provided input and output 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):

$$RSS = \sum_{i=1}^{n} (y_i - x_i \hat{\theta})^2$$

Where x_i ***denotes*** i^{th} ***row (*** i^{th} ***data sample) in the input data matrix X, and*** $\hat{\theta}$ ***is a column vector.***

Task 2.3:

Compute the log-likelihood function for every candidate model:

$$\ln P(D|\hat{\theta}) = \frac{n}{2}\ln(2\pi) - \frac{n}{2} ln(\hat{\sigma}^2) - \frac{1}{2\hat{\sigma}^2} RSS$$

Here, $\hat{\sigma}^2$ ***is the variance of a model's residuals (prediction errors) distributions*** $\hat{\sigma}^2 = \frac{RSS}{(n-1)}$***, with n the number of data samples. D denotes the input-output dataset (X, y)***

Task 2.4:

Compute the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) for every candidate model:

$$AIC = 2k - 2 \ln p\,(D|\hat{\theta})$$

$$BIC = k \ln(n) - 2 \ln p\,(D|\hat{\theta})$$

Here, $\ln p\,(D|\hat{\theta})$ ***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 variable (x5) is subject to 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 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 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.*

Submission Instructions

Requirement	Details
File Naming	NAME_studentID
File Format	.docx/.pdf format
Submission Method	Campus 4.0 platform (submission link provided 2 weeks before deadline)

Marking and Feedback

How will my assignment be marked?

Your assignment will be marked by the Module Team using standardized criteria.

How will I receive grades and feedback?

Provisional marks will be released once internally moderated. Feedback will be provided alongside grades release within 2 weeks (10 working days).

What will I be marked against?

Details of the marking criteria for this task can be found in the Assessment Marking Criteria section at the end of this brief.

Grade Requirements

You must achieve 40% or above to pass this assessment. Ensure you understand the marking criteria for successful completion.

Assignment Support and Academic Integrity

Getting Help

If you have any questions about this assignment, please meet your respective module leader or teacher for more information.

Language Standards

You are expected to use effective, accurate, and appropriate language within this assessment task.

Academic Integrity

The work you submit must be your own. All sources of information need to be acknowledged and attributed; therefore, you must provide references for all sources of information and acknowledge any tools used in the production of your work, **excluding Artificial Intelligence (AI)**.

We use detection software and make routine checks for evidence of academic misconduct. Definitions of academic misconduct, including plagiarism, self-plagiarism, and collusion can be found in Student handbook in Campus 4.0.

All cases of suspected academic misconduct are referred to for investigation, the outcomes of which can have profound consequences to your studies.

Support for Students with Disabilities

If you have a disability, long-term health condition, specific learning difference, mental health diagnosis or symptoms, contact the Student Support Office for assistance.

Unable to Submit on Time?

If events prevent you from submitting on time, guidance on extenuating circumstances is available in the Student Handbook or from the Student Support Office.

Administration of Assessment

Module Leader Name:	Hikmat Saud
Module Leader Email:	stw0032@softwarica.edu.np
Assignment Category:	Written
Attempt Type:	Standard
Component Code:	CW

Assessment Criteria

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 of figures and tables.	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.