STA 3010: GROUPS AND GROUP WORK ASSIGNMENT

Group 1

1. Nzioka, Trizah Mbithe,
2. Tanui, Rovy Kiplang'at,
3. Waiya, Collins Gitau
4. Emmanuel

Group 2

1. Mbagaya, Alexia Nyandiko,
2. Iranzi, Innocent,
3. Queen Esther Kibegi

Group 3

1. Oyabi, Sammi Makana
2. Gesicho, Rose_Delilah Nyanganyi
3. Omenge, Nathan Orang'o

Group 4

1. Magu, Rita Wawira
2. Elue, Chiadika Somto
3. Mwangi, Calvin Gacheru

Group 5

1. Kariuki, Kyra_Joy Wawira
2. Mohamed, Mohamed Abdulaziz_
3. Othum, Wambui, Esther

Group 6

1. Murithi, Fiona Kinya
2. Barigye, Cedric Francis
3. Amondi, Margaret Thatcher
4. Abdullahi, Abdirisak Hussein

Group 7

1. Omar, Tanveer, Nyambane
2. Mark Chweya
3. Lavendar Nchagua Mathias

Group 8

1. Korir, Kevin Kipkoech
2. Otieno, Susan Awino
3. Mise, Selmah Tzindori

Group 9

1. Kahsay, Ambachow Ykalom
2. Malachi, Jennifer John
3. Collins Gishangi

Group 10

1. Kuot, Yar Deng_Machar
2. Muema, Julie Koki,
3. Mwasaghua, Moses Elvis

Group 11

1. Sekoh, Hope Neema_Akinyi
2. Muhia, Wilson_Junior Wambugu
3. Muthaura, Mitchel Makena

Group 12

1. Hans Onesmo
2. Patrick Kariuki
3. Mark Mayaki

ASSIGNMENT DESCRIPTION

Course: Statistical Modelling (Year 3)

Weight: 20 Marks

Due Date: _**17TH October 2025**_

Objective

The purpose of this assignment is to help students understand regularization and dimension-reduction techniques in regression analysis. Students will learn how and why ridge regression, lasso regression, and partial least squares (PLS) regression are applied in practice.

Tasks

1. Research & Theory (20%)

Provide a concise literature review of:

- Ridge Regression: motivation, mathematical formulation, penalty term.
- Lasso Regression: motivation, mathematical formulation, feature selection ability.
- Partial Least Squares (PLS): concept, difference from Principal Component Regression, when it is preferred

Include all the references used

2. Data Acquisition (20%)

- Identify suitable dataset with at least 10 predictor variables (continuous or categorical) with a continuous response variable.
- Provide a short description of the dataset: source, variables, sample size, and why the data is appropriate for this analysis.

3. Model Fitting & Analysis (40%)

Fit the three models to your dataset using R and generate reports on the coefficients, model selection criteria and diagnostic plots.

4. Interpretation & Discussion (20%)

Compare the results of the three methods; Which predictors are retained/removed in LASSO? How do Ridge and PLS handle multicollinearity differently? Which model gives the

best predictive performance? Provide a clear conclusion with practical implications based on your dataset.

Deliverables

1. Report (max 10 pages):

- Introduction & Motivation
- Methodology (formulas and explanation)
- Results (tables and plots)
- Discussion & Conclusion
- References

2. R Script File (.R) or RMarkdown (.Rmd) with all code.

3. Dataset file (if publicly available, include link or upload).