University of the Philippines Cebu
College of Science
Department of Computer Science
CMSC 178DA (Special Topics)
Introduction to Data Analytics

ACTIVITY GUIDE (WEEK 12)

Activity 4
Applying Multiple Linear Regression Analysis to Multivariate Data

Objective

The activity for this module aims to demonstrate practical knowledge in performing multiple linear regression analysis technique, with an added dimension of incorporating a categorical variable to the multiple linear regression model, as well as how to interpret the multiple linear regression output, e.g., R^2, model parameters, among others.

Tasks

1. Partner with the same classmate as in the previous activity (unless if you want to change partner, with permission from me) since this activity is to be accomplished in pairs to promote collaboration and interaction. (*In case the class size is odd-numbered, one student is to volunteer to work individually.*)

2. Read-analyze together the given case study (adapted from R. Jain, 1991) below:
   *Let us reuse the performance data on remote procedure call (**RPC**) mechanism for two (2) mainframe operating systems (OS) - UNIX and ARGUS – in Activity 3. The performance metric was total elapsed time (in milliseconds), which was measured for various data sizes (see Table 1). Fit a multiple linear regression model that combine the sample measurements on total elapsed time in processing with various data sizes for the two OSes.*

   Table 1. **Measured RPC Times on UNIX and ARGUS**.

UNIX		ARGUS	
Data Bytes	**Time**	**Data Bytes**	**Time**
64	26.4	92	32.8
64	26.4	92	34.2
64	26.4	92	32.4
64	26.2	92	34.4
234	33.8	348	41.4
590	41.6	604	51.2
846	50.0	860	76.0
1060	48.4	1074	80.8
1082	49.0	1074	79.8
1088	42.0	1088	58.6
1088	41.8	1088	57.6
1088	41.8	1088	59.8
1088	42.0	1088	57.4

   Source: R. Jain (1991).

   The data analyst poses this question to himself: ***Can we reasonably predict the total elapsed time of RPCs based on the various processed data sizes and type of OS (i.e., UNIX and ARGUS)?***

3. Discuss with your partner the given case above. Agree on the responses required by this activity. Active participation by each member of the pair during the discussion *cum* interaction is expected.

4. Create a multiple linear regression model with the following specification: $\boldsymbol{y = b_0 + b_1 x_1 + b_2 x_2}$, where $\boldsymbol{y}$ = elapsed time, $\mathbf{x_1}$ = data size, and $\mathbf{x_2}$ = OS type (with ARGUS as the reference).

5. Use either of the following application software tools: MS Excel (if it can be done), R, or Python, (or a mix of it) to accomplish task #4.

6. List the steps your pair has undertaken and/or provide the codes (for either R or Python, whichever is applicable) in accomplishing Task #4. Present the narrative in an outline format. Provide appropriate headings/comments above the codes for clarity.

7. Based on your pair's interpretation on the multiple linear regression output, answer the following questions:
   7.1. Is the fitted multiple regression model statistically valid, or meaningful? Justify.
   7.2. How much variation of the total elapsed time can be explained by the multiple regression model? Does this answer the question posed by the data analyst? Justify.
   7.3. Which model parameter(s) contributed meaningfully in predicting the total elapsed time for RPCs? Identify. Justify.
   7.4. What is the per-byte processing cost (i.e., elapsed time) on the UNIX OS? Identify. Justify.
   7.5. What is the final multiple linear regression model specification which you can now use in predicting the processing elapsed times for RPCs? Specify.

8. Name the PDF file of your pair's submission with the following file-naming convention: **Class Activity [number]-[Your Family Name]_[Your Partner's Family Name]**. For example, "Class Activity **4**-**Lao_Uy**". The highlighted portion of the example are the ones you supply information for your pair.

9. Structure the document for your responses following required sections below:
   9.1. Cover Page (format accordingly)
      9.1.1. Course Number
      9.1.2. Class Activity Number and Title
      9.1.3. 'Submitted by' information (pair members, course section, and schedule)
   9.2. The Case (in the next page)
   9.3. The Multiple Linear Regression Output
      9.3.1. Fitted Model Summary Output
      9.3.2. Narrative Text (i.e., Task#4, *in outline format*)
   9.4. Pair's interpretation (*refer back to Task 7 for context).*

10. Submit a copy of your pair's Activity 4 responses (in PDF format) to your respective 'Class Activity' submission bin of the Learning Management System (LMS). In certain instances when your Internet connectivity is unstable, you may submit via email, but you must inform me ahead (i.e., not after the deadline).

Assessment Criteria

- **Fitted Model Summary Output - 40%**
  (*correctness in generating appropriate multiple linear regression analysis summaries*)
- **Narrative Text – 20%**
  - ✓ Fitting multiple linear regression model (10%)
  - ✓ Data-encoding for categorical variable (10%)
- **Interpretation – 30%**
  (*analytical accuracy on observations, descriptions, and implications*)
- **Packaging Submission – 10%**
  (*shows creativity in organizing and presenting the submitted document which contains the pair's responses*)