MAST7866 Foundations of Data Science

Group Report Assessment 2025

One member of your group should submit your report in PDF electronically using the Turnitin link on the MAST7866 Moodle page by **Friday 28nd November 2025**, clearly listing the names of all group members.

Group members will all be awarded the **same** mark unless there is a clear discrepancy in contribution.

The report must include an **appendix** containing **documented** R code for the work presented, **both** for statistical analyses and for graphics.

Task: Linear regression analysis of the following dataset relating to medical insurance costs. You aim to examine how well you can **predict** the **annual_medical_cost** using the other variables. You need to pre-process (for example, certain transformations) and analyse the whole dataset and to build a predictive multiple linear regression model for the annual medical cost based on the training sub-dataset, diagnosis of your model assumptions, and evaluate the generalisability of your model on the testing sub-datasets by calculating mean predicting square errors. The analysis should include some visualization approaches to provide insights about the data distribution and relationships.

Dataset: The dataset is partially adopted and modified from a medical database. It contains demographic details along with medical history and insurance details. The dataset can be splitting into the training and testing sets by using the memberships indicated in the last column of the data spread sheet.

The files GroupX.csv (on Moodle, where X is your group number) contains 54 variables： **person_id， age， sex， region, urban_rural, income， education， marital_status。 employment_status， household_size， dependents， bmi (**Body Mass Index**)， smoker， alcohol_freq， visits_last_year ， hospitalizations_last_3yrs， days_hospitalized_last_3yrs， medication_count， systolic_bp， diastolic_bp， ldl (**low-density lipoprotein**)， hba1c， plan_type， network_tier， deductible， copay (**a contribution made by an insured person towards the cost of medical treatment**)， policy_term_years, policy_changes_last_2yrs, provider_quality, risk_score, annual_medical_cost, annual_premium, monthly_premium, claims_count, avg_claim_amount, total_claims_paid, chronic_count, hypertension, diabetes, asthma, copd (**Chronic obstructive pulmonary disease**), cardiovascular_disease, cancer_history, kidney_disease, liver_disease, arthritis, mental_health, proc_imaging_count, proc_surgery_count, proc_physio_count, proc_consult_count, proc_lab_count, is_high_risk, had_major_procedure.** Finally, the variable **tnts** is 0 for the observations you should build (train) your model on and is 1 for the observations you should use to test your model. Each group has different observations randomly selected for training and testing.

Examine how well you can **predict** the **annual_medical_cost** using the other variables.

Introduction

Describe the types of variables in the dataset (for example, numerical or categorical, continuous or discrete) and explain what will follow in your report. [10% of marks]

Exploratory Analysis through Descriptive statistics and Graphical summaries

Explore the full set of data (training and test data sets together) using descriptive statistics and graphs to allow the reader of the report to engage with the data. [30% of marks]

Statistical Modelling

Clearly define what linear models you are going to build, using the training data only. Ensure that you justify the use of particular models using diagnosis tools. [20% of marks]

Results and Conclusions

Clearly display the results from your analyses of the training data only and state what conclusions can be drawn. Discuss the results. Can the **annual_medical_cost** be predicted accurately for the test data? How can you improve the accuracy of prediction [30% of marks]

Appendix of R code

Tidy R script annotated with comments to explain each part of the process. [10% of marks]

Page limit

Reports should be a maximum of 20 pages in length (including all tables and figures, but excluding the appendix of R code).