

REVIEW Assignment 1 - Regression Models

- Due 31 Mar 2023 by 23:59
- Points 100
- Submitting a file upload

Applied DS for Innovation

Assignment 1

The Brief:

In this assignment, we will tackle a regression problem. We will be working on a dataset consolidated from census data in the USA. The goal is to accurately predict cancer mortality based on information related to US counties. The dataset contains 33 different features (demography, medical information).




The metric used to assess model performance is MSE (mean squared error).

Assignment:

The assignment is composed of 4 different parts with specific scope and constraint:

- Part A: Experiment on univariate linear regression. You are allowed to pick only 2 features from the dataset and will train an univariate linear regression for each of them.
- Part B: Experiment on multivariate linear regression. You are allowed to use all numeric features from the dataset and train a multivariate linear regression model.
- Part C: Experiment on multivariate linear regression with feature engineering or with any algorithms of your choice.
- Part D: Final report detailing the different steps of the projects, results achieved, issues faced and recommendations. Use the steps defined by the CRISP-DM methodology for the sections of your report.

Submission Requirements:

- Part A to C: [Experiment reports](https://docs.google.com/document/d/16zUW7dJJ5qH3Wpn3U0i2sXFngQMsXnQLcfm3JEszOXk/edit?usp=share_link)  (https://docs.google.com/document/d/16zUW7dJJ5qH3Wpn3U0i2sXFngQMsXnQLcfm3JEszOXk/edit?usp=share_link) (in PDF or Word) based on the provided template and corresponding [Jupyter notebook](https://colab.research.google.com/#scrollTo=-Rh3-Vt9Nev9)  (<https://colab.research.google.com/#scrollTo=-Rh3-Vt9Nev9>).
- Part D: A final report detailing this project following the [CRISP-DM methodology](https://www.datascience-pm.com/crisp-dm-2/)  (<https://www.datascience-pm.com/crisp-dm-2/>). The report should not exceed 1500 words.

All assignments need to be submitted before the due date on Canvas. Penalties will be applied for late submission.

Assessment Criteria:

- Quality of data exploration (visual + summary stats)
- Strength of justification for features selected and model used
- Quality of code and accuracy of results
- Appropriateness of the CRISP-DM framework usage
- Depth of discussion of ethics/privacy issues, value, benefits and recommendation for business

Dataset:

- Training set: [cancer_us_county-training.csv](https://drive.google.com/file/d/1qe5KgJrITsw1h8t-fJhqJw2DMtcfaW5/view?usp=share_link)  (https://drive.google.com/file/d/1qe5KgJrITsw1h8t-fJhqJw2DMtcfaW5/view?usp=share_link)
- Testing set: [cancer_us_county-testing.csv](https://drive.google.com/file/d/1OT6Y7TXT4c630XPKrl-uh3DmcSOpOypq/view?usp=share_link)  (https://drive.google.com/file/d/1OT6Y7TXT4c630XPKrl-uh3DmcSOpOypq/view?usp=share_link)

IMPORTANT NOTE:

You need to use your UTS email to access the data.

Data Dictionary:

TARGET_deathRate: Dependent variable. Mean *per capita* (100,000) cancer mortalities(*a*)

avgAnnCount: Mean number of reported cases of cancer diagnosed annually(*a*)

avgDeathsPerYear: Mean number of reported mortalities due to cancer(*a*)

incidenceRate: Mean *per capita* (100,000) cancer diagnoses(*a*)

medianIncome: Median income per county (*b*)

popEst2015: Population of county (*b*)

povertyPercent: Percent of populace in poverty (*b*)

studyPerCap: *Per capita* number of cancer-related clinical trials per county (*a*)

binnedInc: Median income per capita binned by decile (*b*)

MedianAge: Median age of county residents (*b*)

MedianAgeMale: Median age of male county residents (*b*)

MedianAgeFemale: Median age of female county residents (*b*)

Geography: County name (*b*)

AvgHouseholdSize: Mean household size of county (*b*)

PercentMarried: Percent of county residents who are married (*b*)

PctNoHS18_24: Percent of county residents ages 18-24 highest education attained: less than high school (b)

PctHS18_24: Percent of county residents ages 18-24 highest education attained: high school diploma (b)

PctSomeCol18_24: Percent of county residents ages 18-24 highest education attained: some college (b)

PctBachDeg18_24: Percent of county residents ages 18-24 highest education attained: bachelor's degree (b)

PctHS25_Over: Percent of county residents ages 25 and over highest education attained: high school diploma (b)

PctBachDeg25_Over: Percent of county residents ages 25 and over highest education attained: bachelor's degree (b)

PctEmployed16_Over: Percent of county residents ages 16 and over employed (b)

PctUnemployed16_Over: Percent of county residents ages 16 and over unemployed (b)

PctPrivateCoverage: Percent of county residents with private health coverage (b)

PctPrivateCoverageAlone: Percent of county residents with private health coverage alone (no public assistance) (b)

PctEmpPrivCoverage: Percent of county residents with employee-provided private health coverage (b)

PctPublicCoverage: Percent of county residents with government-provided health coverage (b)

PctPublicCoverageAlone: Percent of county residents with government-provided health coverage alone (b)

PctWhite: Percent of county residents who identify as White (b)

PctBlack: Percent of county residents who identify as Black (b)

PctAsian: Percent of county residents who identify as Asian (b)

PctOtherRace: Percent of county residents who identify in a category which is not White, Black, or Asian (b)

PctMarriedHouseholds: Percent of married households (b)

BirthRate: Number of live births relative to number of women in county (b)

(a): years 2010-2016

(b): 2013 Census Estimates

Template:

Experiment Report: [link](#) 

(https://docs.google.com/document/d/16zUW7dJJ5qH3Wpn3U0i2sXFngQMsXnQLcfm3JEszOXk/edit?usp=share_link).