P8130: Biostatistical Methods I

Final Project (Fall 2018)

Due, December 17th @ 1:00pm

Guidelines for Project Submission

This group project must be submitted through CourseWorks before the deadline. Email submissions WILL NOT be accepted and will receive a score of 'Zero' for all group members!!

All graphs, output and interpretations must be included in **ONE PDF** (not the R file), otherwise it will not be graded. In a separate attachment, you also have to submit the R code used in your.

General Writing Instructions

Your project should not exceed **5 double-spaced pages** using 11 or 12-point font, EXCLUDING figures and tables, references, appendix, that can be placed at the end of the five summary pages. Be selective in your output and visual displays!

Your report should be structured as a publishable research article containing the following sections:

- Abstract
- Introduction (context, background of the problem)
- Methods (data description and statistical methods)
- Results
- Conclusions/Discussion

Your findings should be written as for an informed (but non-statistical) audience (**no formulae!**). Each figure and table should be of publishable quality and well notated, i.e., labeled and/or captioned.

Grading Instructions

The rubric attached will be used to evaluate the project. This is a group project and collaborations within your group are essential and great practice for your career.

Academic dishonesty will be punished with a 'Zero' grade for this project.

The data for this project were aggregated from multiple source including American Community Survey census.gov, clinicaltrials.gov, and cancer.gov. The dataset includes several variables per county (USA) and your task is to build a regression model that 'best' predicts cancer mortality (target_DeathRate). You can use county data and/or group by state; also feel free to use only variables (a), (b), or a combination of the two. You are most welcome to consider other alternatives, as long as you provide a logical justification.

Aspects that need to be addressed in your report:

- Data exploration: descriptive and visualization
- Re-code/combine levels of categorical variables based on frequency and practical importance
- Model diagnostics
    - Heteroscedasticity, normality and multicollinearity
    - Functional form for continuous predictors
- Outliers and missing values
- Predictive capability of the model
- In this course, we only covered linear regression models. Let us assume that even after exploring different combinations of predictors your model does not fit the data well and/or does not have a good predictive ability. What other statistical methods/models not covered in this course would you recommend for future steps?

<u>‘Cancer_Registry.csv’ Dictionary:</u>

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 population 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*)

PctPubliceCoverageAlone: 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