Statistical Learning for Public Policy I

Problem Set 2

Oct 12, 2022

This lab uses data from Fatehkia et al. (2019). Their paper investigates the predictability of crime rates across urban neighborhoods with the help of interests recorded on Facebook. They use the Facebook Advertising API to collect data on different interests of Facebook users of a particular ZIP code area. They combine the collected user data with demographic information. Then, they fit various statistical models to predict assault, burglary, and robbery rates as a function of the Facebook and census demographic variables.

A subset of the authors' data are stored in `fcbk.csv`. The table below provides a description for each variable in the dataset. All variables with `acs_` are from the American Community Census (2015) and all variables with `fb_` are from Facebook.

Name	Description
`zip`	ZIP code
`state`	State
`city`	City name
`assaults17`	Total reported assaults in 2017 per 100,000 capita
`acs_pop15_19`	% of population aged 15-19
`acs_pop18_24`	% of population aged 18-24
`acs_med_age`	Median age
`acs_white`	% of population one race White
`acs_black`	% of population one race Black or African-American
`acs_foodstamp`	% of households on food stamp benefits
`acs_med_inc`	Median family income
`acs_inc150k`	% households with income > 150K
`acs_inc25k`	% households with income <= 25K
`acs_pop18_24_nohs`	% of population 18-24 with less than high school degree
`acs_pop18_25_college`	% of population 18-24 with bachelors or higher degree
`acs_pop25_nohs`	% of population 25+ with less than high school degree
`acs_pop25_college`	% of population 25+ with bachelors or higher degree
`fb_hip_hop_music`	% of Facebook users (older than 18) interested in Hip Hop
`fb_[...]`	% of Facebook users (older than 18) interested in ...

Before working on this problem set, you might consider working through the Linear Regression Lab in ISL (p. 109-114).

Question Set 1

a) How many observations are in the data? How many variables are there?

b) What share of the total number of observations is within the state of California?

c) In which zip codes are the most (the least) assaults per capita?

Question Set 2

Fit two bivariate linear regressions: In model A we wish to predict assaults using information on the share of Facebook users interested in Hip Hop music. In model B we wish to predict assaults using information on the share of Facebook users interested in first-person shooter games.

a) Which of the two predictors accounts for more variation in the outcome variable?

b) Predict the assault rate for a zip code in which the share of Facebook users interested in in first-person shooter games or hip hop music is equal to the average in the data. Construct a confidence interval for your predicted value.

c) Construct a model C that leverages both the share of Facebook users interested in Hip Hop music and users interested in first-person shooter games to predict assaults. What is the increase in R^2 relative to the best-performing bivariate linear model?

d) Suppose that the share of Facebook users interested in Hip Hop music increases from 20 percent to 30 percent, i.e., a 10 percentage point increase. How many more assaults should we expected according to model C? What about an increase from 50 to 60 percent?

Question Set 3

Construct a linear regression model that uses all continuous variables in the dataset.

a) What is the increase in R^2 relative to the best-performing bivariate linear model?

b) Predict the assault rate for a typical zip code (i.e., a zip code for which all variables are centered on their mean) and compute the standard error. What percent of the uncertainty for the predicted value is due to estimation uncertainty?