INFO370 Problem Set: Logistic Regression, Prediction

August 1, 2021

Instructions

This PS has the following goals:

1. learn to use and interpret logistic regression results
2. learn to handle categorical variables
3. learn to predict the outcomes, and compare with the actual data.

1 Who will win the elections? (50pt)

This question asks you to do a simple election model. We are looking for the U.S. 2020 presidential elections by counties. Your task is to model the winner (1/0 for democrats winning the presidential elections in this county), and explain the winner using population density, minority population, education level, income, and geographic differences (the census region).

The data file is called *us-elections_2000-2020.csv.bz2*. The variables are

FIPS county FIPS code

year election year

state state name

state2 2-letter state code

region census region (west, midwest, south, northeast)

county county name

office President (we look only at presidential elections)

candidate name of the candidate

party party of the candidate

candidatevotes votes received by this candidate for this particular party

totalvotes total number of votes cast in this county-year

income personal income, USD/per capita (BEA data)

population population, census estimate (BEA data)

LND010200D land area (sq.mi) at 2000 (Census data)

EDU600209D Persons 25 years and over, total 2005-2009

EDU695209D Educational attainment - persons 25 years and over - bachelor's degree 2005-2009

POP010210D Resident population (April 1 - complete count) 2010

POP220210D Population of one race - White alone 2010 (complete count)

POP250210D Population of one race - Black or African American alone 2010 (complete count)

POP320210D Population of one race - Asian alone 2010 (complete count)

POP400210D Hispanic or Latino population 2010 (complete count)

PST110209D Resident total population estimate, net change - April 1, 2000 to July 1, 2009

The complex variable names originate from the US Census.

1. (2pt) Load data, and do basic sanity checks.
2. (5pt) You are going to work with 2020 data. However, some important information for 2020 is missing. Fill the missings with the most recent values that exist in the data.

   Hint: check out `DataFrame.fillna` method. Ensure you order your observations right and do not fill missings with values from other counties!
3. (13pt) Make a new data frame that only contains 2020 data, and that contains a binary variable: the county went to democrats in 2020.

   Note: you have to build that variable using two lines of data in the original data frame *by FIPS* after the data is ordered by year.
4. (12pt) Create auxiliary variables: population density (population divided by land area); minority percentage; and percentage of college graduates. These can be made of different variables, and as none of these are changing fast, it should not have much of an impact.

   Ensure the variables are in a reasonable range!

   Hint: there are values that do not make sense. Remove those.
5. (8pt) Estimate logistic regression model where you explain democrats' vote share with population density, minority percentage, education level, income, and census region.
6. (10pt) Interpret the results. Which results are statistically significant?

   Note: you may want to change some of the units, e.g. you may want to measure population density in 1000/per sq mi, instead of persons per sq mi.

2 Predict AirBnB Price (50pt)

Your next task is to analyze the Beijing AirBnB listing price (variable *price*). It is downloaded from Inside Airbnb but we suggest to use the version on canvas (*airbnb-beijing-listings.csv*). You have to work with several sorts of categorical variables, including those that contain way too many too small categories. You are also asked to do log-transforms, interpret the results, and do some predictions.

1. (2pt) Load the data. Select only relevant variables you need below. Even better, check out the `usecols` argument for `read_csv`. Do basic sanity checks.
2. (4pt) Do the basic data cleaning:

(a) convert *price* to numeric.

(b) remove entries with missing or invalid price, bedrooms, and other variables you need below

3. (5pt) Analyze the distribution of *price*. Does it look like normal? Does it look like something else? Does it suggest you should do a log-transformation?

   Hint: consult lecture notes Section 4.1.7 *Interactions and Feature Transformations*.

4. (5pt) Convert the number of bedrooms into another variable with a limited number of categories only, such as 1, 2, 3, 4+, and use these categories in the models below.

   Hint: consult the python companion for lecture notes `http://faculty.washington.edu/otoomet/machinelearning-py/cleaning-data.html#cleaning-data-converting-variables`

5. (6pt) Run an OLS (i.e. fit the linear regression model) where you explain the price with number of bedrooms where bedrooms uses these four categories. Interpret the results, including R^2.

   Hint: if 0-BR is the reference category, the effect for 1BR should be -12.62 (but it may depend on how exactly do you clean data).

   R^2 is explained in lecture notes 4.1.3 "Model evaluation: MSE, RMSE, R^2".

6. (8pt) Now repeat the process with the model where you analyze log price instead of price. Interpret the results. Which model behaves better in the sense of R^2?

   For the following tasks use either *log(price)* or *price*, depending on your answer here.

7. (5pt) Include further two variables into the model: *room type* and *accommodates*. While room type only contains three values, *accommodates* contains many different categories. Recode these as "1", "2", "3", "4 and more".

   Fit this model. Interpret and comment the more interesting/important results. Do not forget to explain what are the relevant reference categories and R^2.

8. (5pt) Now use the model above to predict (log) price for each listing in your data.

9. (5pt) Compute root-mean-squared-error (RMSE) of this prediction.

   RMSE is explained in lecture notes, 4.1.3 "Model evaluation: MSE, RMSE, R^2".

10. (5pt) Now use your model to predict the price for a 2-bedroom apartment that accommodates 4 (i.e. a full 2BR apartment).