

JSC270 Winter 2021 Assignment 2

Exploring income and its correlated factors in 1994 America


In 1995 Kohavi and Becker extracted and cleaned [this data](#) from the 1994 US Census. It went on to be used as a benchmark dataset for many papers advancing statistical and machine learning methods. Here we will work with this dataset to explore patterns of income in the population and consider how time and location may alter or confound these relationships.

This assignment is to be completed individually. To submit this assignment, upload a zip file formatted *studentnumber_assig2.zip* with 3 files included:

1. A pdf of the answers to the questions below with a link to your github repo that includes the python notebook with the solutions.
2. A 2-page pdf report (described below).

Github. This assignment can be done in google colab or on your local machine. (1 pts):


1. Fork [this repo](#) to your Github account. Here are some straight-forward instructions from stackoverflow if you need them:



I think that the "most polite way" to do so would be:

7

1. Fork the original repo on your GitHub account
2. Checkout a new branch for your changes `git checkout -b <your_branch_name>` (in case you didn't do that before)
3. Add a new remote for your local repository: `git remote add github <your_repository_ssh_url>`
4. Push your beautiful new branch to your github repository: `git push github <your_branch_name>`



In this way you will have a repo forked to the original one, with your changes committed in a separate branch. This way will be easier in case you want to submit a pull request to the original repo.

Share Improve this answer Follow

edited Apr 3 '17 at 18:19

answered Aug 13 '13 at 3:53



llekn

2,163 1 15 21

-
- 1 Took me a minute to note that `github` in these instructions is the bit usually referred to as `origin`. Otherwise, straightforward and simple. – leanne Jan 5 '17 at 17:39

From: <https://stackoverflow.com/questions/18200248/cloning-a-repo-from-someone-elses-github-and-pushing-it-to-a-repo-on-my-github>

2. Invite your TAs Matt (12mre1) and Lauren (larunerdman) to be collaborators of the repo from your Github account.

3. Create a (colaboratory or other python) notebook and link it to your github repo in your new branch. Include a link to colaboratory.
4. Copy the starter code you need from what was initially in the github repo and get going!

Initial data exploration. (2 pt):

1. Check the columns of your data. Are they the expected data types based on their descriptions in [this text file description of the data](#)?
2. How are missing values represented in this data? Cast any misspecified variables to a different data type or assert that they are and set missing characters to np.nan. Count the number of missing values in each column.
3. Individually plot the distributions of *capital_gain* and *capital_loss*. Should these variables be transformed? Why/why not? If yes, create a new variable for either that should be transformed and plot or describe in a table the distribution of the new categorical variable.
4. This data is census data and therefore used to estimate finite population means using sampling. The weights here indicate the share of the population that sample represents based on location and, sometimes, other factors. You can read more about the weights in the paper linked in [this text file description of the data](#). Plot or numerically explore the distribution of *fnlwgt* Is this data symmetrically distributed? Compare the distribution of this variable between men and women and comment on any trends you notice. Should outliers be excluded? If you think yes, set the *fnlwgt* values for those you deem to be outliers as missing for the remainder of your analyses.

Correlation. Use correlation tests or plots which show correlation to answer the following questions (4 pts):

1. Find the correlations between *age*, *education_num*, and *hours_per_week*.
 - a. Do any of the variables appear to be correlated?
 - b. Statistically test any variable pairs with a correlation coefficient $> |0.1|$ for its difference from 0 and report your result. Is the direction and significance of your finding as expected?
 - c. How does the correlation (and its significance) between *education_num* and *age* compare between male and female participants? Is this expected?
 - d. Compare the weighted vs unweighted variance and covariance between *education_num* and *hours_per_week*. What do the changes in these values tell us about the weights in our data and who may be over/under represented?

Regression. Using linear regression, answer the following questions (4 pts):

1. Fit a linear regression with hours_per_week as the dependent variable and sex as the independent variable.
 - a. Do men tend to work more hours?
 - b. Add education_num as a control variable, does the trend in hours worked by men vs women remain the same? Is education_num significant?
 - c. Now add gross_income_group as a binary variable in the model and compare the model with these 3 variables vs the models with 2 and 1 variable. What statistic would you choose to decide which model is the best? Describe how you could re-do what was just done using a model fitting procedure.

Report (4 pts): Write a (max 2 page) report providing a brief background, motivation, methods, results, and discussion of one of the questions asked in this assignment. Make sure it is accessible for someone who doesn't already know you are working with this data set.

Presentation (Separate from assignment, due Feb 11: 5 pts): Create a 2 minute recorded presentation providing a brief background, motivation, methods, results, and discussion of the topic you wrote your report on.

Bonus question (1 pt): The estimator of regression slope coefficient in a univariate regression can be expressed as:

$$\hat{\beta} = \frac{\text{cov}(X, Y)}{\text{var}(X)}$$

How does this coefficient mathematically relate to a correlation coefficient between X and Y? Answers can be submitted in any format (latex, word, image of written math) included in the pdf answers for this homework.