Lab 6: Statistical Exploratory Data Analysis

In lab 5 we developed expertise with hypothesis testing by implementing a few tests from scratch. Our goal in lab 6 is to use our new-found expertise with hypothesis testing to conduct an exploratory data analysis. To do this we will be considering life expectancy again, but this time we will be working with a new data set of US county level information. As usual, our goal will be to work with a response variable and use the techniques we have developed so far this term to explore which variables might make good predictors for this response.

1 Read in and prepare the data

1. First read the counties.csv file into your python development environment using pandas. You'll need to import pandas and then apply the '.read_csv' method.

2. We will also do some basic feature engineering here by making five new variables.

   (a) Divide the number of covid deaths in each county in 2022-03-01 by the 2019 population – add a column called covid-deaths_total_per_capita containing this data to the dataframe.

   (b) Divide the number of covid confirmed cases in each county in 2022-03-01 by the 2019 population – add a column called covid-confirmed_total_per_capita containing this data to the dataframe.

   (c) Create an indicator that measures whether a county has greater-than-average life expectancy. Add a column called above_average_life-expectancy containing this data to the dataframe.

   (d) Record for each county the largest industry by number of employees. You will need to use the 20 variables in the data named industry/.../employees. To do this, you will need to fill missing values in these variables. For simplicity, fill missing values with 0 (you may find .fillna useful). Put this information into a single column called biggest_industry.

   (e) Record for each county the modal educational level. You will need to use the 4 variables in the data named edu/.... Put this information into a single column called county_modal_ed.

3. Answer the following questions in your writeup.

   (a) When we constructed the above_average_life-expectancy indicator was it appropriate to use the mean for this purpose?

(b) When we constructed the above_average_life-expectancy indicator how could you have used hypothesis testing to build this variable? If you had, how many categories would it have had?

(c) What are we implicitly assuming by filling missing values with 0 when we create the largest industry variable?

2 Life expectancy

1. In this section we are going to explore which variables are predictive of US county level life expectancy. Thus we will be using life-expectancy as our dependent variable. We will focus on the following variables as potential predictors:

```
X = ['state'
        ,'longitude (deg)'
        ,'latitude (deg)'
        ,'noaa/temp'
        ,'noaa/altitude'
        ,'male'
        ,'deaths/suicides'
        ,'deaths/homicides'
        ,'bls/2020/unemployed'
        ,'avg_income'
        ,'covid-deaths_total_per_capita'    #constructed
        ,'covid-confirmed_total_per_capita'    #constructed
        ,'covid-vaccination/2021-12-01'
        ,'county_modal_ed'    #constructed
        ,'poverty-rate'
        ,'cost-of-living/living_wage'
        ,'cost-of-living/food_costs'
        ,'cost-of-living/medical_costs'
        ,'cost-of-living/housing_costs'
        ,'cost-of-living/tax_costs'
        ,'health/Average Number of Mentally Unhealthy Days'
        ,'health/% Smokers'
        ,'health/% Adults with Obesity'
        ,'health/% Physically Inactive'
        ,'health/% Long Commute - Drives Alone'
        ,'biggest_industry']    #constructed
```

2. For each numerical potential predictor variable, use scipy.stats.linregress() to estimate the Pearson's correlation coefficient and the statistical significance (p-value) of the correlation against the life-expectancy variable, e.g. scipy.stats.linregress(df["life-expectancy"], df["avg_income"]).

3. We can test for association between categorical and numerical variables using a Kruskal-Wallis test via the scipy.kruskal() function. In this example, we want to know if the distribution of life expectancy differs by state:

```
samples_by_group = []
for value in set(data["state"]):
    mask = data["state"] == value
    samples_by_group.append(data['life-expectancy'][mask])

stat, p = kruskal(*samples_by_group)
```

4. In a single table, indicate the variable name, test statistic, p-value, and whether there is a statistically significant relationship between that variable and life-expectancy at a threshold of $\alpha = 0.05$, using all the hypothesis tests conducted in 2.2 and 2.3.

5. Answer the following in your writeup:

   (a) Explain what it means to reject the null hypothesis in a Pearson's correlation coefficient hypothesis test.

   (b) Discuss the meaning of Pearson's correlation coefficient values obtained in your analysis. How does it help in interpreting the relationship between the variables under study?

   (c) For which variables did the obtained correlation coefficient align with your initial expectations? Why or why not?

   (d) Reflect on the assumptions of Pearson's correlation vs Spearman's correlation. How would you verify these assumptions in your analysis?

   (e) Explain what it means to reject the null hypothesis in a Kruskal-Wallis hypothesis test.

   (f) Explain why the Kruskal-Wallis test was chosen as an appropriate non-parametric alternative for comparing more than two independent groups.

   (g) Regarding a p-value, fixing a threshold of 0.05 means that we are setting the risk of rejecting the null hypothesis when we shouldn't at 5% (this is a type I error). In this section you've conducted over 20 hypothesis tests. What may this imply about the likelihood you've made at least one type I error?

3 Classification on above average life expectancy

1. In this section we are going to explore which variables are predictive of US county level life expectancy. Thus we will be using above_average_life-expectancy as our dependent variable.

2. As above, run a Kruskal-Wallis test for each numerical variable versus the above_average_life-expectancy indicator.

3. We can test two categorical variables for association using a χ^2 (read chi-squared) test of independence. The "normal" χ^2 goodness-of-fit test tests if one set of categorical counts was generated from the same distribution as a second set of categorical counts. The test can also be used to test for independence of two variables.[1] To use the normal χ^2 goodness of fit test to check independence, expected frequencies of co-occurrences of the values from the two variables are calculated under the assumption that the values are independent. The χ^2 test is then used to determine if the co-occurrence counts of the other data set match the expected independent distribution (null hypothesis). If the counts do not match, then you reject the null hypothesis of independence. The χ^2

[1] Loosely speaking, two variables are independent if they have no relationship.

test formulated for checking independence is available in scipy as the chi2_contingency() function:

```
combination_counts = pd.crosstab(data['state'], data['above_average_life-expectancy'])
chi2, p, _, _ = chi2_contingency(combination_counts)
```

Run a χ^2 test of independence between each categorical variable versus the above_average_life-expectancy indicator.

4. In a single table, indicate the variable name, test statistic, p-value, and whether there is a statistically significant relationship between that variable and above_average_life-expectancy at a threshold of $\alpha = 0.05$.

5. Answer the following in your writeup:

   (a) What does it mean to reject the null hypothesis in a χ^2 test of independence?

   (b) Reflect on the limitations of hypothesis testing methods in general. In what ways do these methods provide insight into data, and what are their potential drawbacks or pitfalls?

   (c) You have dealt with two versions of the response variable – life expectancy in years (a numerical variable) and an indicator showing which counties have above average life expectancy (a binary categorical variable). Did the results of your hypothesis tests depend on which of these you used? That is, were there any predictors that had statistically significant relationships with one of these two versions of the response but not the other? How would this affect future machine learning?

   (d) Reflect on any challenges or insights gained during the process of conducting hypothesis tests and interpreting the results. What lessons have you learned that could inform future research or data analysis endeavors?

4 Submission Instructions

- Write up the answers to the questions in a short word document; aim for around 2 pages of text and include all graphics generated. Add footnotes identifying which sentence addresses which questions. Write in complete sentences organized into paragraphs – your goal is to explain what you've done and what you've learned to your audience (me!). Accordingly, you may seek to emulate some of the sections of the whale paper describing their data. Include the appropriate plots you've generated as mentioned above. Convert this to pdf and submit it. Submit your .ipynb file as well.

- The grading rubric for this assignment will be available in Canvas.

- NO OTHER SUBMISSION TYPES WILL BE ACCEPTED.

- **Late policy**: 5% of total points deducted per day for three days – after that no submissions allowed.