Data Analytics Report

Sushma G Herakal
PES1UG19CS528
PES UNIVERSITY
Bangalore, Karnataka
sushmagh92@gmail.com

Tejashwini Hosamani
PES1UG19CS539
PES UNIVERSITY
Bangalore, Karnataka
onlyteju5@gmail.com

Introduction & Background

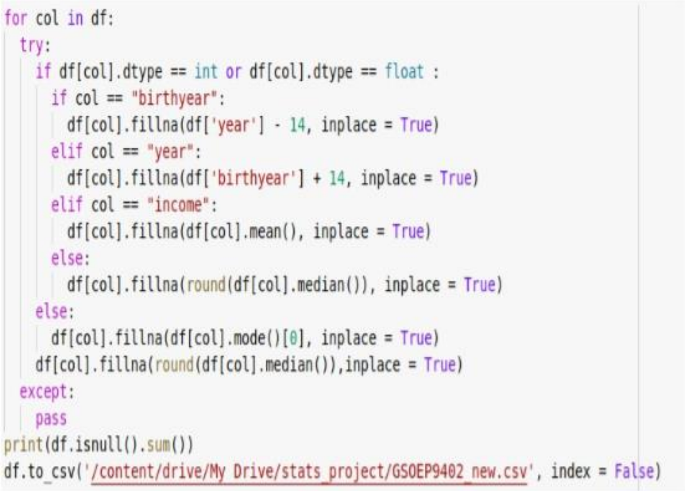
Our dataset is a cross section data for 675 14-year-old children born between 1980 and 1988. The sample is taken from the German Socio - Economic Panel (GSOEP) for the years 1994 to 2002 to investigate the determinants of secondary school choice. After cleaning up our dataset, our main analysis was based on the variations of the various continuous, discrete and categorical variables with the school choice of the children. We also tried to test the mean income of the population, the linear relationship between the variables and inferences were drawn from the results produced.

This sample from the German Socio-Economic Panel (GSOEP) for the years between 1994 and 2002 has been selected by Winkelmann and Bose (2009) to investigate the determinants of secondary school choice. In the German schooling system, students are separated relatively early into different school types, depending on their ability as perceived by the teachers after four years of primary school. After that, around the age of ten, students are placed into one of three types of secondary school: "Hauptschule" (lower secondary school), "Real Schule" (middle secondary school), or "Gymnasium" (upper secondary school). Only a degree from the latter type of school (called Abitur) provides direct access to universities. A frequent criticism of this system is that the tracking takes place too early, and that it cements inequalities in education across generations. Although the secondary school choice is based on the teachers' recommendations, it is typically also influenced by the parents; both indirectly through their own educational level and directly through influence on the teachers. The validity of family background variables instrumenting education in income regressions has been much criticized. In this paper, we use data from the 2004 German Socio-Economic Panel and Bayesian analysis to analyse to what degree violations of the strict validity assumption affect the estimation results.

We show that, in case of moderate direct effects of the instrument on the dependent variable, the results do not deviate much from the benchmark case of no such effect (perfect validity of the instrument's exclusion restriction). In many cases, the size of the bias is smaller than the width of the 95% posterior interval for the effect of education on income. Thus, a violation of the strict validity assumption does not necessarily lead to results which are strongly different from those of the strict validity case. This finding provides confidence in the use of family background variables as instruments in income regressions.

Previous work

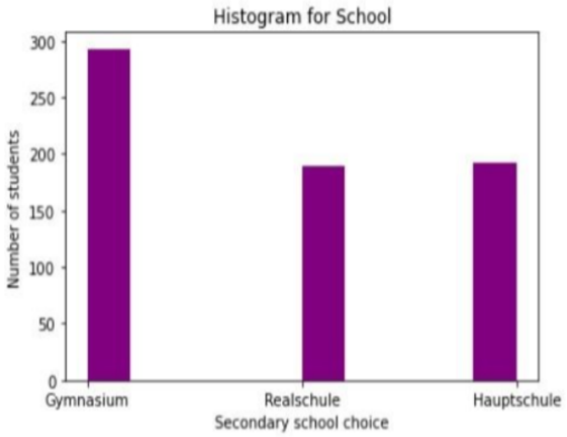
There were around 20 missing values in each variable. Our dataset had 3.11% missing values. We replaced the missing values in the variable ‘income’ with the mean of the column. For the variables birth year and year, from the description of the dataset we knew that the children under consideration were 14 year olds, in order to calculate the missing birth year, 14 was deducted from the year. If the year was missing then 14 was added to the birth year. In case, both year and birth year were missing, we replaced the values with the median. The other numerical missing values were replaced with the median of the variable and rounded off to the nearest integer. The missing values in categorical variables were replaced with mode of the variable. Fixing inconsistent capitalization. Only the variable ‘state’ had inconsistent capitalisation. All the words were first converted into lower and then the title method was used to fix the capitalization.



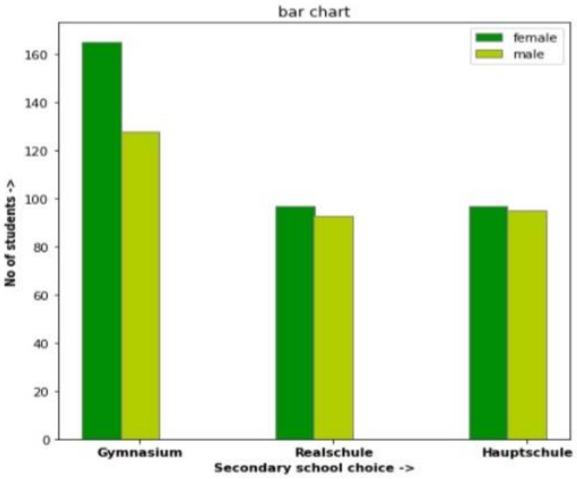
```
for col in df:
  try:
    if df[col].dtype == int or df[col].dtype == float :
      if col == "birthyear":
        df[col].fillna(df['year'] - 14, inplace = True)
      elif col == "year":
        df[col].fillna(df['birthyear'] + 14, inplace = True)
      elif col == "income":
        df[col].fillna(df[col].mean(), inplace = True)
      else:
        df[col].fillna(round(df[col].median()), inplace = True)
    else:
      df[col].fillna(df[col].mode()[0], inplace = True)
    df[col].fillna(round(df[col].median()),inplace = True)
  except:
    pass
print(df.isnull().sum())
df.to_csv('/content/drive/My Drive/stats_project/GSOEP9402_new.csv', index = False)
```

Normalization : The aim of Normalization is to change the values in the dataset to a common scale, without changing the differences in the range of values. Normalization only changes the numeric values of mean, median and mode without affecting the overall behaviour and distribution of the dataset. There are many Normalization techniques: The maximum absolute scaling, The min-max feature scaling, Z-score method. The first part of data analysis involved graph visualisation. We plotted 3 graphs

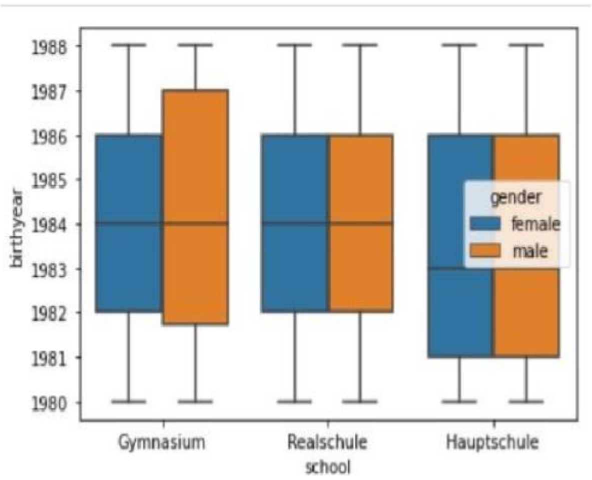
1. Histogram
2. Bar chart
3. Boxplot



Histogram was plotted for the variable ‘school’. In order to determine the popularity of the schools among students, a histogram was plotted with school as the variable. The inferences drawn were: Most of the students, (approximately 300 of them) opted for Gymnasium as their secondary school choice. Real Schule and Hauptschule were equally popular among students.



A bar chart was plotted to compare the number of female and male students in each school. The inferences drawn were: The number of female students in gymnasium(around 165) is more than the male students(around 125). While they are almost equally distributed in Real Schule and Hauptschule.



In order to find out about the spread of values of birth year among the various schools with respect to gender, a box plot was constructed. The inferences drawn were: Real Schule and Gymnasium have most of their students born in 1984, whereas Hauptschule students were born in 1983. The spread of data for male students is much higher than female implying that the popularity of this school was more consistent among males.

Components & its Explanation:

Variable Description Age of individual in years East-German 1 if individual resides in East-Germany; 0 otherwise Foreigner 1 if individual is a foreigner; 0 otherwise Urban 1 if individual resides in a city with more than 100,000 inhabitants; 0 otherwise Income 700-999 Euro 1 if household equivalent income is between 700 and 999 Euro; 0 otherwise Income 1,000-1,299 Euro 1 if household equivalent income is between 1,000 and 1,299 Euro; 0 otherwise Income more than 1,300 Euro 1 if household equivalent income is more than 1,300 Euro; 0 otherwise Education High school degree 1 if individual has a high secondary school degree; 0 otherwise Intermediate school degree 1 if individual has an intermediate secondary school degree; 0 otherwise Basic school degree 1 if individual has a basic school degree; 0 otherwise Student high school 1 if individual still attends a high school; 0 otherwise Student intermediate school 1 if individual still attends an intermediate school; 0 otherwise Student basic school 1 if individual still attends a basic school; 0 otherwise Student other school 1 if individual still attends another kind of school including kind of school unknown; 0 otherwise Job Full-time job 1 if individual has a full-time job including civil-/military service; 0 otherwise Part-time job 1 if individual has a part-time job; 0 otherwise In vocational training 1 if individual is in vocational training (not student); 0 otherwise Unemployed 1 if individual is unemployed and looking for a job; 0 otherwise Parents Mother high school degree 1 if mother has a high school degree: 0 otherwise Mother intermediate school degree 1 if mother has an intermediate secondary school degree: 0 otherwise Mother basic school degree 1 if mother has a basic school degree: 0 otherwise Father high school degree 1 if father has a high school degree: 0 otherwise Father intermediate school degree 1 if father has an intermediate secondary school degree: 0 otherwise Father basic school degree 1 if father has a basic school degree: 0 otherwise Marital Status Living together with partner 1 if individual is married or single, but lives together with her partner; 0 otherwise Not living together with partner 1 if individual is single and does not live together with a partner; 0 otherwise Single, but living together un- 1 if individual is single, but no information known if individual lives together with her partner.

Experimental Results

With accuracy of 0.85 model has reached its best.

From the histogram we concluded that most of the students opted for Gymnasium as their secondary school choice. By comparing the number of female and male students in each school, it was observed that the number of female students is more than the male students in gymnasium, while they were almost equally distributed in the other schools. This analysis helped us understand how the popularity of schools varied through the years in Germany. The correlation coefficients are scatter plots were used to conclude that, most of the variables were independent. The variables year and birth year have a very strong positive correlation of 1.There is positive correlation between the size of the family and number of kids. There is a positive correlation between parity and number of kids. The hypothesis test concludes, it is plausible that the mean income of the population may be greater than or equal to 70000.

In Germany, the development of smoking behaviour among youths gives rise to worry, because smoking participation rates among youth steadily increased during the decade up to 2003. Moreover, individuals of later cohorts tend to start smoking earlier than older ones. When aiming at lower smoking participation rates special focus should lie on the decision of young individuals to start, since it might be easier to prevent individuals from starting the habit than to make smokers stop smoking. This paper contributes to the existing literature by providing an econometric analysis on smoking onset using contemporary information at the time individuals started as well as retrospective information and thereby allowing for a more causal interpretation of the results. Furthermore, the analysis is done for a country where price changes of cigarettes are unlikely to have caused the increase in smoking incidence rates.

Conclusion

Contribution : Sushma G Herakal & Tejashwini Hosamani has equally contributed in this project.

References :

- Bertram, L., A. Beckhoff, I. Demuth, S. Dyzel, R. Eckardt, S.-C. Li, U. Lindenberger, G. Pawelec, T. Siedler, G.G. Wagner, E. Steinhagen-Thiessen (2014), Cohort Profile: The Berlin Aging Study II (BASE-II). International Journal of Epidemiology 43: 703–712.Search in Google Scholar
- Deaton, A. (2015), Nobel Prize Lecture by Angus Deaton. Available under :Search in Google Scholar
- Reinecker, Ph., K. Erhardt, M. Kroh, P. Transverter (2017), The Request for Record Linkage in the IAB -SOEP Migration Sample. SOEP Survey Papers 291, .Search in Google Scholar
- Federal Ministry of Labour and Social Affairs (2017), The German Federal Government's 5th Report on Poverty and Wealth. .Search in Google Scholar
- Federal Press Office (2016), Government Report on wellbeing in Germany. Berlin. Search in Google Scholar