**Part 1: Data Preparation**

1. Load Data Set

Ein Bild, das Tisch enthält.

Automatisch generierte Beschreibung

1. Descriptive Statistics

Ein Bild, das Tisch enthält.

Automatisch generierte Beschreibung

1. Adjusted Summary Statistics + Missing/Implausible Values?

Ein Bild, das Tisch enthält.

Automatisch generierte Beschreibung

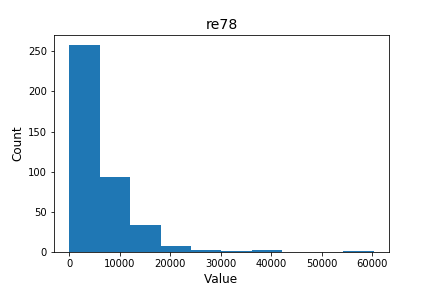
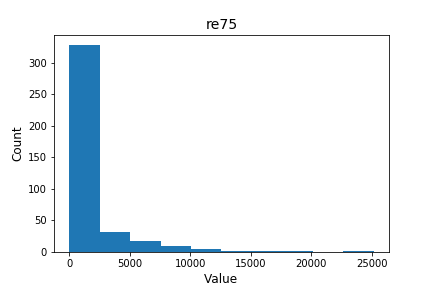
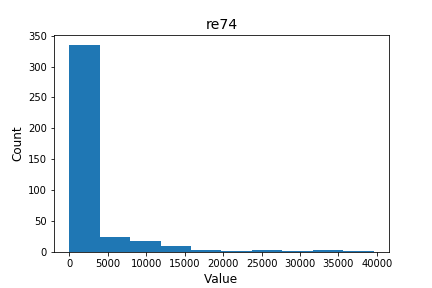
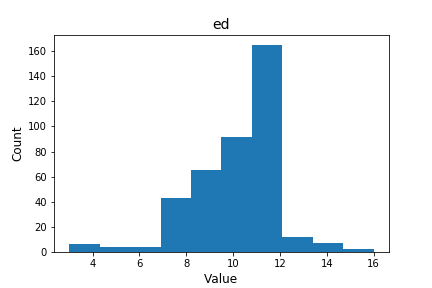
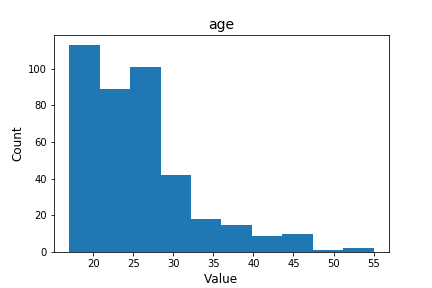
﻿According to the summary statistics, the data set does not contain any missing values, which can be seen from the missing values equal to zero and total observations sum up to 400 for all included variables. Furthermore, the dummy variables are all correctly coded. This can be concluded as the min value shows 0, the max value shows 1 and the number of unique observations is 2 for all dummy variables. With a focus on the continuous variables, the “age” shows plausible values ranging from 17 to 55. Additionally, the annual earnings in 1974, 1975 and 1978 also shows plausibility while ranging from 0-﻿39’570.68, 0-﻿25’142.24, and 0-﻿60’307.93, respectively. Lastly, the mean, variance and standard distribution seems plausible for all variables, which finally indicates full plausibility across the data set.

1. Column Drop of “age2”

Ein Bild, das Tisch enthält.

Automatisch generierte Beschreibung

1. Histograms of Continuous Variables



According to the histogram of the variable “age”, the sample is positively skewed which leads to a higher density of younger people compared to older people. Furthermore, depending on the model definition, ages above 45 could be defined as outliers and hence be eliminated from the data set. Based on the histogram of the variable “ed”, the education in years in the sample is slightly negatively skewed and suggests that the majority of people in the sample has not attended any further education (like university) after high school. Lastly, according to the histograms for “re74”, “re75” and “re78”, the earnings in the years 1974, 1975 and 1978 in the sample are strongly positively skewed, which underlines traditional earnings distribution in history. While the majority has earnings approximately between 0 and 15’000 USD, only a few people in the sample earned more, up to 35’000, 25’000 and 60’000, respectively. Such a strong skewness can lead to major issues in analytics processes as outlying observation can have a strong effect on the analytics target, e.g., predictors and its coefficients. In order to prevent such effects, either statistical transformations could be run (e.g., log-transformations) or an outlier threshold could be defined to eliminate outlying observations based on the earnings.

The proportions of the treatment and control group in the data set can be concluded from the dummy variable “treat”. The mean value of 0.41 shows that 41% of the people in the sample were assignment in the treatment group, while 59% of the people in the sample were used as the control group. Hence, the treatment assignment was not balanced for this experiment.

1. Data Anomalies + Potential Issues for Statistical Analysis

According to the comparison of the distribution of the ages and earnings in the sample, the positive and negative skewness are in accordance. Due to a higher proportion of younger people, it is plausible that the earnings have a higher proportion in lower earnings and that the positive skewness becomes less over the three observations years as more people in the sample are either eligible to work or have more experience which could lead to higher earnings. Furthermore, the small proportion of people with higher degree educations (beyond high school) is also in accordance with the small proportion of high earnings in the sample, as higher education degrees can be associated with higher earnings. Hence, as a conclusion, no data anomalies can be detected in the data set.

Apart from that, the skewness in the variables still determines a major issue, why the data in its raw form cannot be used for statistical analysis. As mentioned in part e), the skewness needs to be corrected by either a data transformation (e.g., log-transformation) or by defining threshold for outliers and their elimination from the data set. These distributions can lead to biases in analytics processes as outlying observations can have a strong effect on the analytics target, e.g., predictors and its coefficients. Furthermore, the degree of randomness across the sample needs to be checked by comparing the treatment and control group and their assigned distributions. For correct statistical analysis, it must be given that the treatment and control assignment is determined independently from the attributes.

1. Balancing Checks

Ein Bild, das Tisch enthält.

Automatisch generierte Beschreibung

As given by the definition of the balancing checks, any standardized differences above 10 are considered as large. According to the table above (measure “StdDiff”), the variables “age”, “ed”, “hisp”, “nodeg” and “re78” are hence determined as large differences between the treatment and control group.

The highest standardized difference can be observed in the degree variable. While 83% of the observations in the control group have no degree, only 73% of the treatment group have no degree. This difference can be stated as highly significant due to a p-value of 0.02 (confidence level = 0.05). The second highest standardized difference can be found in the earnings variable of 1978, where control group has an average income of 4’708.88, while the treatment group has an average income of 6’318.53. Resulting in a standardized difference of 23.27, this result is also highly significant with a p-value of 0.03. Compared to these findings, the variables “age”, “ed” and “hisp”, which have standardized differences of 13.44, 10.06 and 13.80, these results are not statistically significant with p-values of 0.19, 0.33 and 0.17, respectively, with a predetermined confidence level of 0.1. Even higher, but still plausible confidence levels (e.g., 0.1), would still not lead to significance.

As a result, even with five variables that show standardized differences above 10, only “nodeg” and “re78” can be considered as imbalanced covariates across the data set due to their statical significance, while the other covariates can be considered balanced based on the balance checks.