

University of St.Gallen

School of Management, Economics, Law, Social Sciences and International Affairs

Data Analytics II: PC2

University of St. Gallen

Jonas Husmann | 16-610-917

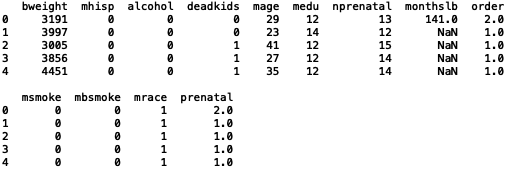
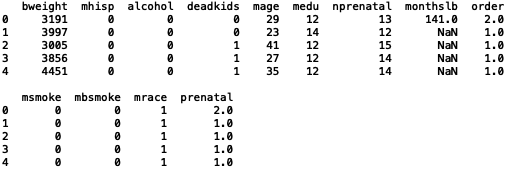
Niklas Leander Kampe | 16-611-618

Prof. Dr. Michael Lechner

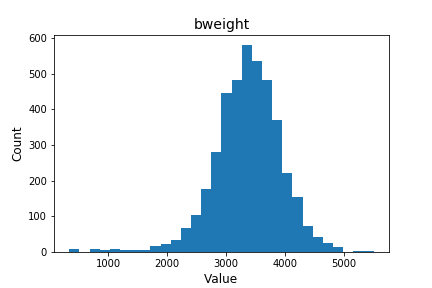
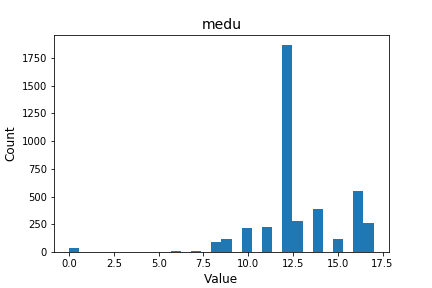
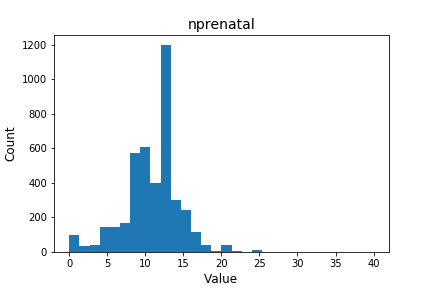
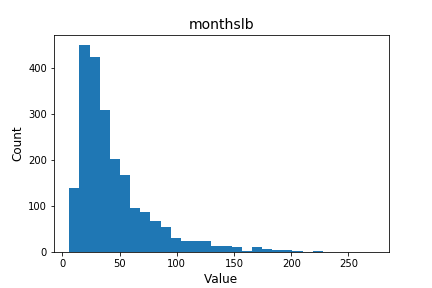
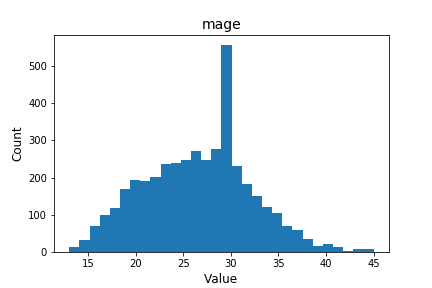
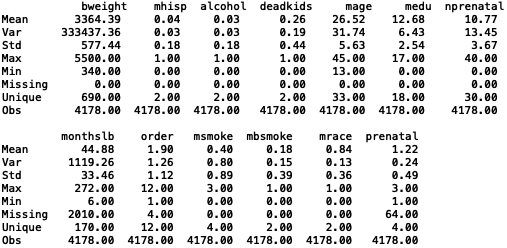
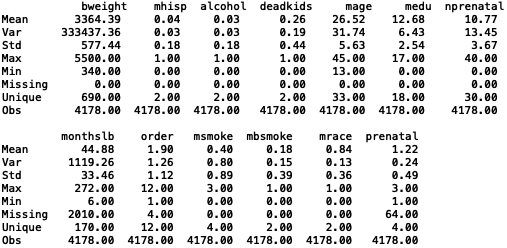
March 07, 2022

**Part 1: Data Preparation**

1. Load Data Set



1. Summary Statistics and Histograms + Anomalies or Missing Values

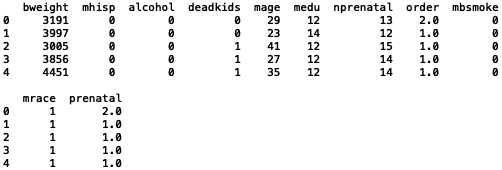
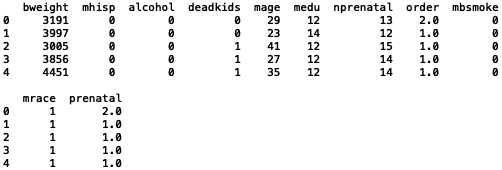


From the summary statistics above, it’s observable that all dummy variables are coded correctly, which can be seen from the min-value of 0, max-value of 1 and the number of unique values of 2. Furthermore, the (continuous) variables “bweights”, “mage”, “medu”, “nprenatal”, “monthslb”, “order”, “msmoke” and “prenatal” seem also to have no anomalies since the min-, max- and number of unique values match the variable definition and their description (e.g., variables coded on a 3-step interval have indeed 3 unique values or reasonable birth weights). These observations are underlined by the histograms. While the target variable “bweight” shows a relatively even distribution with the most weights being between 3000 and 4000 grams (which can be considered as normal for healthy newborns), the covariates “monthslb” and “nprenatal” are slightly positively skewed and the covariates “mage” and “modu” are slightly negatively skewed, but they do not seem to have significant impact for further statistical/econometric analysis. Only potential outliers can be identified, i.e., in “modu”, on which it can be potentially decided if the observations should be omitted. Finally, the only unreasonable value in the data set is given by the min-value of the “monthslb” of 6, which is defined as six months since the birth of the last child, which generally cannot be true. Otherwise, no more data major anomalies can be determined in the data set.

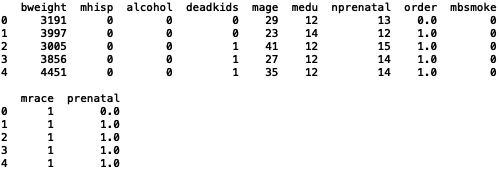
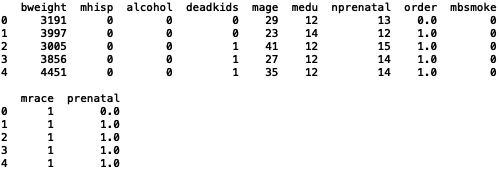
Next to that, the summary statistics identify missing values for the covariates “monthslb”, “order” and “prenatal”. In this context, it is important to mention that the variable “monthslb” has a very significant proportion of missing values, which suggest to rather ignore this covariate instead of omitting the missing observations, due to an otherwise significant reduction of the data set.

Lastly, the mean value of the dummy variable “mbsmoke”, which is defined as the treatment variable, of 0.18 shows that 18% of the sample refers to the treatment group, while the remaining 82% refer to the control group. This suggest a significant treatment-control imbalance, which will be further discussed at a later stage.

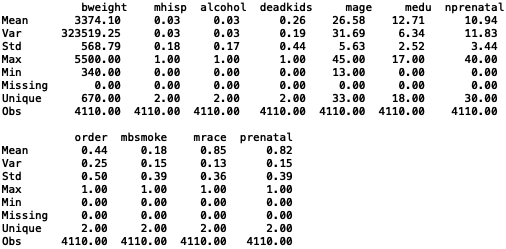
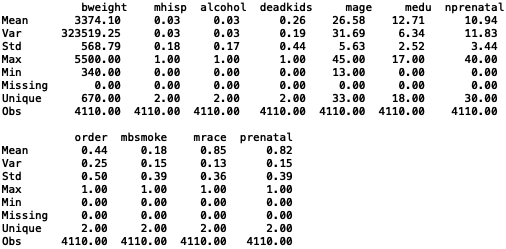
1. Variable and Observation Drops for Missing Values



1. Dummy Variable Recoding



1. Summary Statistics + Suitability for Econometric Analysis



According to the updated summary statistics, the data set does not contain any missing values for all variables. Furthermore, all dummy variables are coded correctly, which can be seen from the min-value of 0, max-value of 1 and the number of unique values of 2. Additionally, all non-dummy variables have non-problematic sample distributions and do also only contain reasonable values, which can be determined based on min-, max- and number of unique values matching the variable definition and their description (e.g., variables coded on a 3-step interval have indeed 3 unique values or reasonable birth weights). Any potential anomalies, which have been identified in task 1a, are also solved. Lastly, the variables “mhisp” and “alcohol” seem to be significantly imbalanced in the whole data set, with only 3% each having a “positive label”. Hence, conclusions from an econometric analysis from these two variables can be difficult based on the distribution and the sample size. Finally, no more major issues can be identified for further econometric analysis.

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 with a predetermined confidence level of 0.05. According to the table above and hence the balancing checks on the covariates (measure “StdDiff”), the variables “alcohol”, “deadkids”, “mage”, “medu”, “nprenatal”, “order” and “prenatal” are determined to have a significant standardized difference between the treatment and control group. Furthermore, all standardized differences of the imbalanced variables are negative, despite of “alcohol” and “deadkids”. This implies that the mean values of the treatment group are lower than those of the control group for all imbalanced variables except for “alcohol” and “deadkids”. The other variables, “mhisp” and “mrace”, can be considered as the only balanced covariates in the data set.

Based on these results, it can also be concluded that the data set includes potential multicollinearity in the covariates, e.g., that women who drink alcohol are also likely to be smoking during pregnancy. Such multicollinearity can have major impacts on statistical and econometric statistics and resulting conclusions, e.g., in biased predictors while estimating the effect of the covariates on the target variable. Additionally, as the only two covariates without imbalanced proportions in the treatment and control group are “mhisp” and “mrace”, the major channels for selections in this sample is likely to be the origin/race of the individuals, which might lead to a selection bias in the sample data set with effects on the degree of randomization.



According to the balance checks on the target variable “bweight”, the large, standardized difference of -47.66 is highly significant with a p-value of 0. Furthermore, the negative sign implies as well that the mean value in the treatment group is lower than the mean value in the control group. Hence, the target variable can be determined as imbalanced between the treatment and the control group in the data set.

**Part 2: Effect Estimation**

1. OLS Function for ATE Estimation without Covariates

A screenshot of a computer

Description automatically generated with low confidence

In the table above we can see the outcome of the OLS without covariates. The estimated ATE is -263.4665 grams, indication that smoking reduces the birthweight by this much on average. This estimated ATE is the same (except for the decimals that are slightly different) as the mean difference that we saw in the balance check in 1f). The underlying identifying assumptions are:

* **Conditional Independence Assumption**: given no covariates included, we assume that no control variable influences both the dependent and independent variable. This is unlikely here and we should include other variables such as *alcohol* (see 1f)
* **Common Support Assumption**: we can consider this to be fine, as we assume no confounding variables
* **Exogeneity of confounders**: assuming no confounding variables, this can be considered fine too
* **Stable Unit Treatment Value Assumption (SUTVA)**: Given that from part 1) we know that there are different intensities of smoking (0-3 scale) this is violated

Given that not all the above mentioned assumptions hold, we cannot be sure that the estimated effect is correct. To design an experiment that would remove the confounding it would be necessary to correct for the differences between the treatment and control group, which would allow to remove the selection bias.

1. OLS Function for ATE Estimation with Covariates

As already touched upon in question 2a) we now correct for the differences between the treatment and control group and include the covariates that had a significant and large (above the 10 treshold defined in 1f) difference based on the balance check results. Therefore, we excluded *mhisp* and *mrace* and this OLS regression. We find that the estimated effect of smoking on birthweight is at -226.6595 grams less negative than it was before (see table below).

A screenshot of a computer

Description automatically generated with low confidence

With regards to the assumptions, we again require the same assumptions to hold as in 2a) that is:

* **Conditional Independence Assumption**: Based on the results from 1f) and the included covariates here we assume that no control variable influences both the dependent and independent variable. There are potential other control variables that are unobserved that potentially jointly influence the outcome and selection (will be revisited below)
* **Common Support Assumption**: we can consider this to be fine, as can be seen in the balance check of 1f)
* **Exogeneity of confounders**: This can be considered valid. However, keeping in mind that smoking could potentially also have an influence on other behaviors (such as drinking alcohol).
* **Stable Unit Treatment Value Assumption (SUTVA)**: Same as in 2a) this is violated given that we know from part 1) that there are different intensities of smoking (0-3 scale).

Further unobservable covariates that could threaten our identification strategy could be:

* The educational level of the father
* The household income level
* Mental health of the mother
* Parent’s birthweight