Positivity Assumption Evaluation

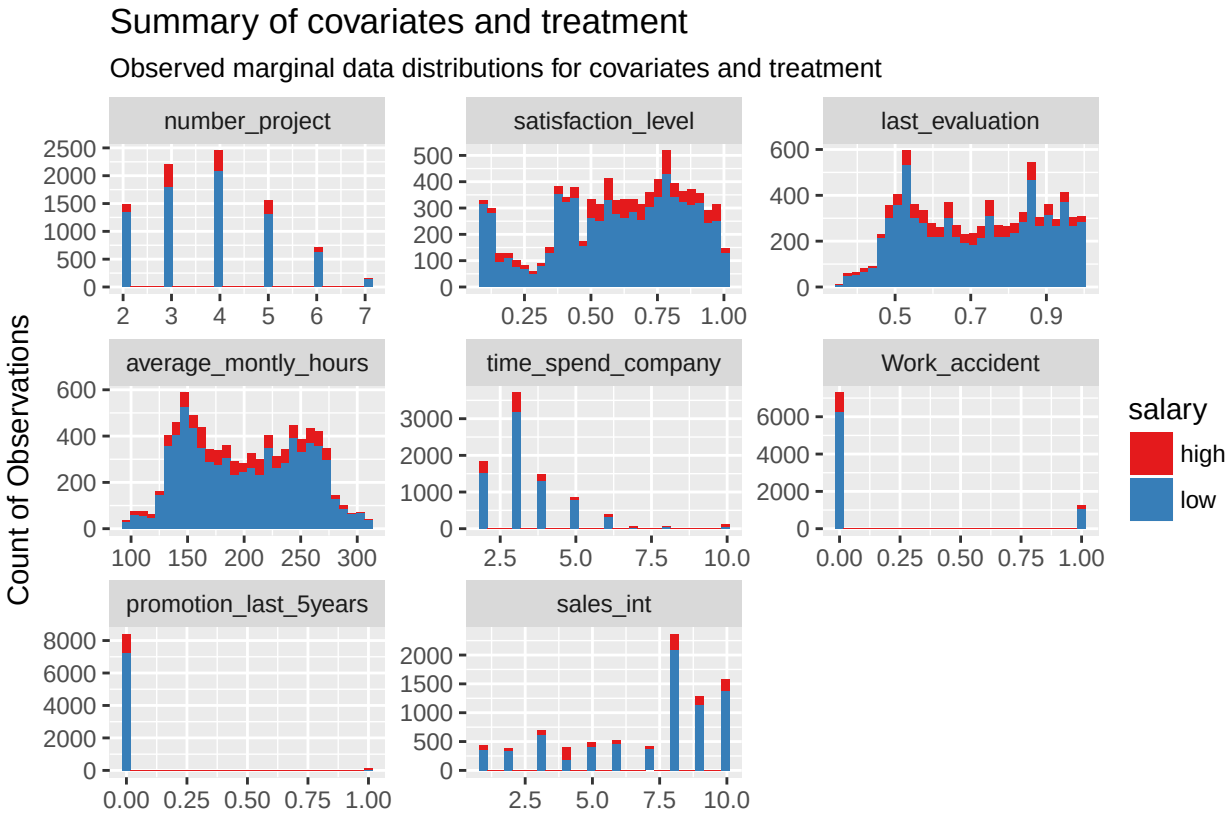
Josiah Davis

4/17/2017

Data Description

There are eight covariates W in the data which are discussed. The goal of this document is to select the number of buckets k to use for approximating the quantitative variables in W. The objective is to maximize the amount of information contained in the approximated variables while removing a minimal amount of observations due to violations of the positivity assumption.

- `satisfaction_level`
- `last_evaluation`
- `number_project`
- `average_montly_hours`
- `time_spend_company`
- `Work_accident`
- `promotion_last_5years`
- `sales`



Observations

- Promotions as well as work accidents are rare at this company.
- Some quantitative variables are discrete (e.g., `number_project`) and some are continuous (e.g., `average_montly_hours`)

Approach

There is one continuous quantitative variable that is not a mediator variable. Quantiles are used to approximate this variable using buckets of size $k = 2, 3, ..., 10$:

- `average_montly_hours`

There are two discrete quantitative variables. Quantiles are used to approximate the variables using buckets of size $k = 2, 3$:

- `number_project`
- `time_spend_company`

There are three variables that are not approximated because they are already binary or becuase they are categories with no natural ordering:

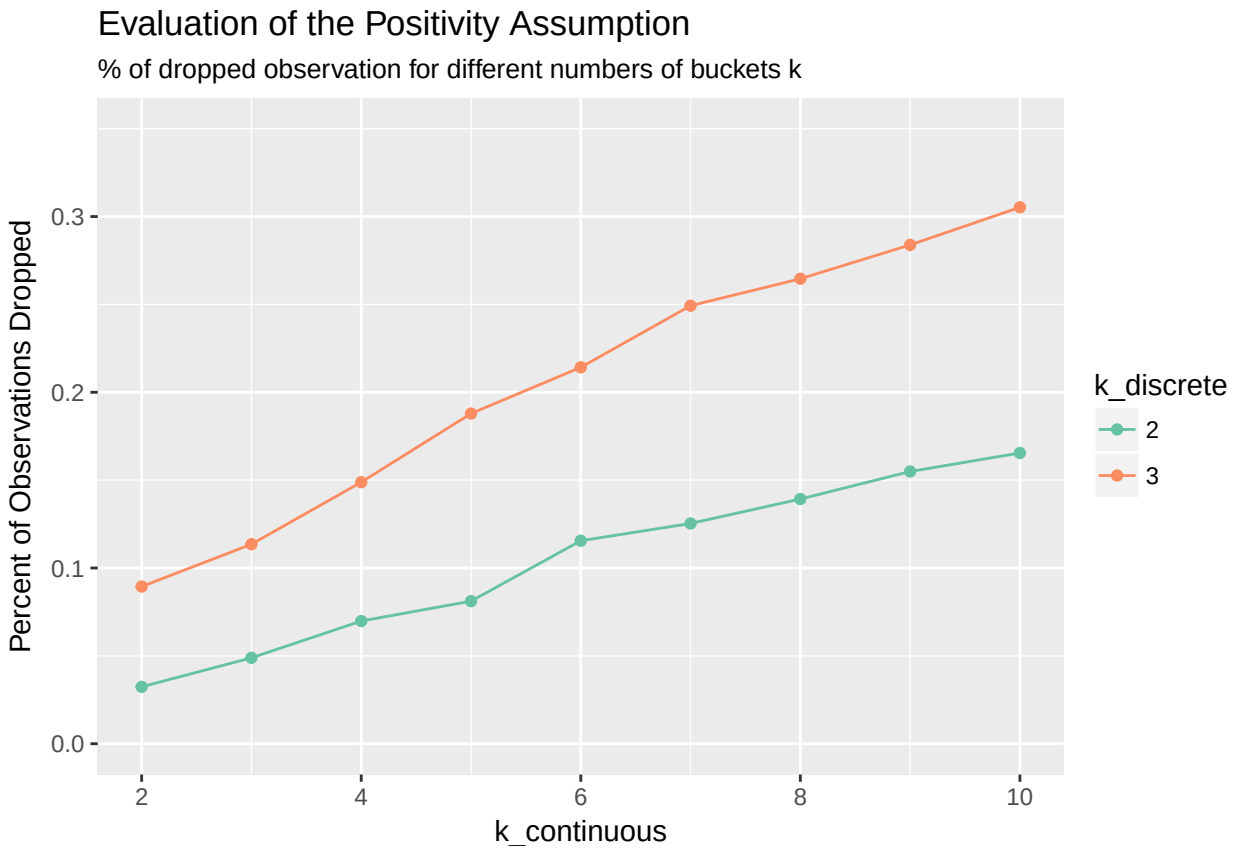
- `Work_accident`
- `promotion_last_5years`
- `sales`

There are three variables that are not approximated because they are mediator variables in our causal graph. This is because we do not need to satisfy the positivity assumption for variables that we are not conditioning on for the G-computation formula.

- `satisfaction_level`
- `last_evaluation`

The percentage of observations that are dropped are calculated for each value fo k for each of the variables.

Results



Observations

- Only **3%** of the observations are lost if quantitative variables are approximated with $k = 2$ buckets
- Approximately **8%** of the observations are lost if continuous variables are approximated with $k = 5$ buckets and quantitative discrete variables are approximated with $k = 2$ buckets.
- Approximately **30%** of the observations are lost if continuous variables are approximated with $k = 10$ buckets and quantitative discrete variables are approximated with $k = 3$ buckets.

Conclusion

So as to miminize the amount of observations lost due to violations of the positivity assumption, while also retaining information from the continuous variables it is recommended that continous variables approimated with no more than $k = 5$ buckets and that discrete quantitative variables are approximated with $k = 2$ buckets.