Positivity Assumption Evaluation

Josiah Davis

4/12/2017

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., `satisfaction_level`)

Approach

There are three continuous quantitative variables. Quantiles are used to approximate the variables using buckets of size $k = 2, 3, ..., 10$:

- `satisfaction_level`
- `last_evaluation`
- `average_montly_hours`

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

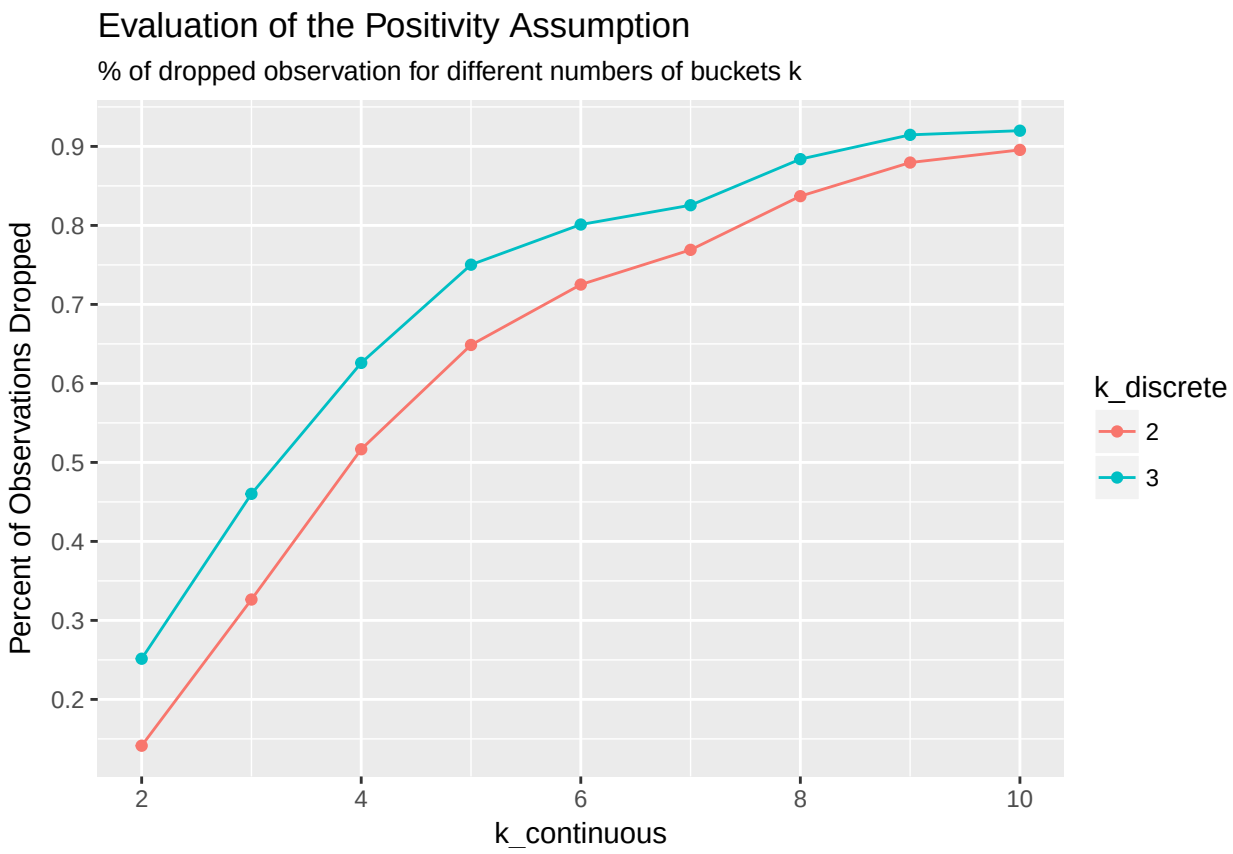
- `number_project`
- `time_spend_company`

There are three variables that are not approximated. This because they are already binary or becuase they have no natural ordering:

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

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

Results



Observations

- Approximately 10% of the observations are lost if using dichotomous approximations for all quantitative variables
- Approximately 33% of the observations are lost if using trichotomous approximations for all continuous variables and dichotomous approximations for quantitative discrete variables
- Approximately 45% of the observations are lost if using trichotomous approximations for all quantitative variables

Conclusion

So as to miminize the amount of observations lost due to violations of the positivity assumption, it is recommended that quantitative variables are transformed into dichotomous approximations ($k = 2$).