Data Analysis Project 1

MA8701

Group 5 : Yellow Submarine

15 February, 2021

In this project, we analyse a real dataset using shrinkage methods from part 1 of the MA8701 course.

Note on Open Science

To pursue the idea of reproducible research, the chosen dataset as well as the code for our analysis are publicly accessible:

- dataset: https://data.ub.uni-muenchen.de/2/1/miete03.asc
- code: https://github.com/FlorianBeiser/MA8701

The Data Set

For our project work we use the Munich Rent 2003 data set as described in https://rdrr.io/cran/LinRegInteractive/man/munichrent03.html. The data set has 12 original covariates, where a brief introduction to these parameters is listed below (in brackets the type of the covariate is explicated), and 2053 observations are available:

- `nmqm`: rent per square meter (double)
- `wfl`: area in square meters (int)
- `rooms`: number of rooms (int)
- `bj`: year of construction (Factor)
- `bez`: district (Factor)
- `wohngut`: quality of location (int)
- `wohnbest`: high quality of location (int)
- `ww0`: hot water supply available (int)
- `zh0`: central heating (int)
- `badkach0`: tiled bathroom (int)
- `badextra`: high-quality bathroom (int)
- `kueche`: upscale kitchen equipment (int) and the response
- `nm`: rental price (double).

"Double" naturally stands for numerical values, "int" categories parameters with integer values, and "Factor" symbolize parameter taking a certain number of levels. The corrplot shows the correlation among the covariates, since `nqnm` has the same effect as `nm`. Therefore, it is excluded from the covariates list for plotting the correlations. It seems that the dataset is suffering a little of the multi-collinearity.

Since the price per square meter `nmqm` multiplied with the area `wfl` directly gives the rental price which the response in the system, it does not make sense to keep both values. Hence, we exclude `nmqm` from the data frame to be able to do serious data analysis on the dataset. The factorial variable `bj` and `bez` introduce 44 and 25 levels, respectively, leading to relatively unclear dependencies between the full covariate set (with factorial variables) and the response, which makes this dataset suitable for a regression analysis and the application of shrinkage methods.

Data Analysis

Subsequently, we start with a plain linear regression model as reference such that we can particularly point out the benefits of shrinkage approaches. As shrinkage methods, we employ the ridge, the lasso and the group lasso. The latter approach seems to be very well suited for our data set, as it allows to take the factorial variables as a single unit for shrinkage.

Regression

We start the data analysis with a vanilla LM regression for reference using R's internal `lm` functionality

The the regression results show a lot of significant covariates. As maybe expected, the area `wfl` is strongly related to the rent price, however confusingly, the significance of different levels of the years of construction `bj` and districts `bez` varies a lot. From those both observations, it is not possible, to extract clear data analysis results, which also would match our interpretation of the problem. (@FB: better formulation)

Ridge Shrinkage

As first shrinkage method, we consider the ridge regression that uses Tikhonov regularisation in the model, where we utilize the `glmnet` library for its implementation in R. Since ridge introduces the additional tuning parameter λ we perform cross validation for the model selection, i.e. for the choice of the optimal λ, where we follow the advice in the ELS to choose λ as the one with minimal CV-error plus one standard deviation of the CV-error.

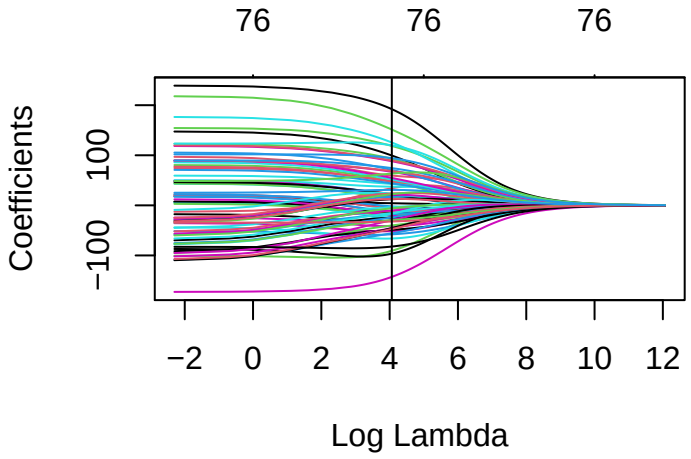


Figure 1: Model selection for ridge shrinkage

In Figure 1, the coefficients for the individual covariates are depicted agains $\log \lambda$, where the optimal λ choice is highlighted. As typical for ridge, the coefficients are shrinked towards 0, but all parameters remain positive weights. This makes the outcome still hard to interpret for our practical data set at hand.

Lasso

In contrast to the previous ridge regression, the lasso adds L_1-regularisation to the regression problem. As before, the implemenation in R relies on the `glmnet` library and the hyperparameter λ is tuned as aforementioned.

In Figure 2, we again see the model coefficients of the covariate set plotted against $\log \lambda$ where the optimal λ chosen by cross validation and the decision rule in ELS in highlighted. Finally and as expected, some coeffiecients are shrinked to 0. However for the optimal λ, some levels of the construction years `bj` and some of the districts `bez` are shrinked to 0 and other would be still significant. For the practical model problem, this is a non-intuitive behaviour.

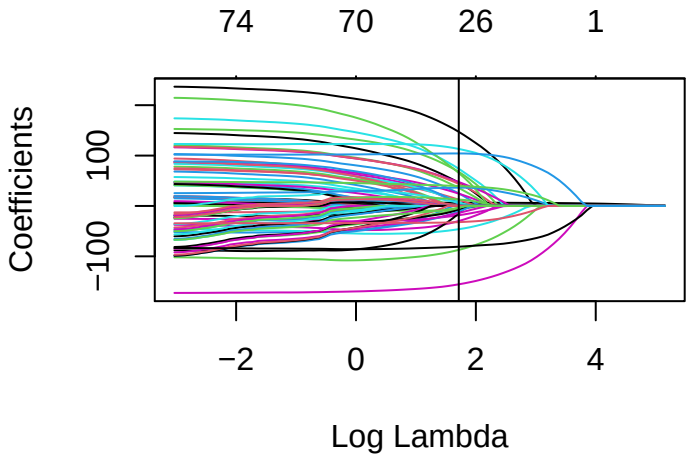


Figure 2: Model selection for lasso shrinkage

Group lasso

The group lasso allows to gather some of the covariates and treat those with the same coefficient jointly in the L_1-regularised problem. For the implementation in R we utilize the `grplasso` library. Naturally, we group the different levels of the factorial variables, i.e. the years of construction `bj` and the different levels for the districts `bez` together, respectively.

Again we employ the same model selection criterion via cross validation as above, but implement it in R using the `gglasso` library. In the cross validation procedure, we can observe sudden jumps when a new group is taken into or out from the model, since the calculatations either include all or non of the levels of the factorial variables, what can yield to a jump in the number of parameters from 9 to 53 in a single step on the lmabda grid.

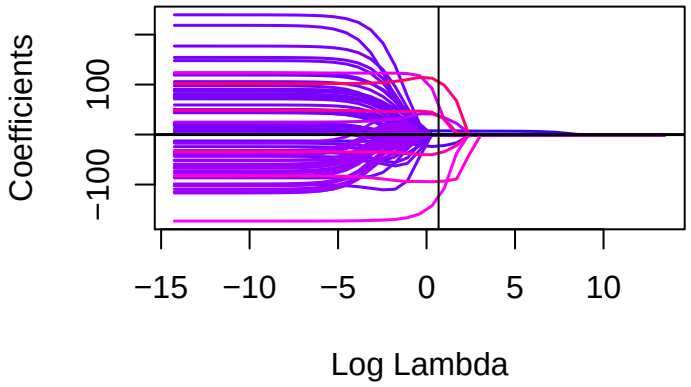


Figure 3: Model selection for group lasso

In Figure 3 we again depict the coefficients of the covariates against the $\log \lambda$ and highlight the optimal hyperparameter. By construction, all levels of a factorial variable are shrinked to simultaneously. This means that either all levels remain in the model or all levels are excluded, what corresponds better to the practical interpretation of those variables.

Conclusion

In this project, we have chosen a practical data set which contains data on rental prices in Munich (two of the group members are practically familiar with the difficult housing situation in Munich and it was appealing to analyse this statistically). For the data analysis, the factorial variables needed special attention. A plain

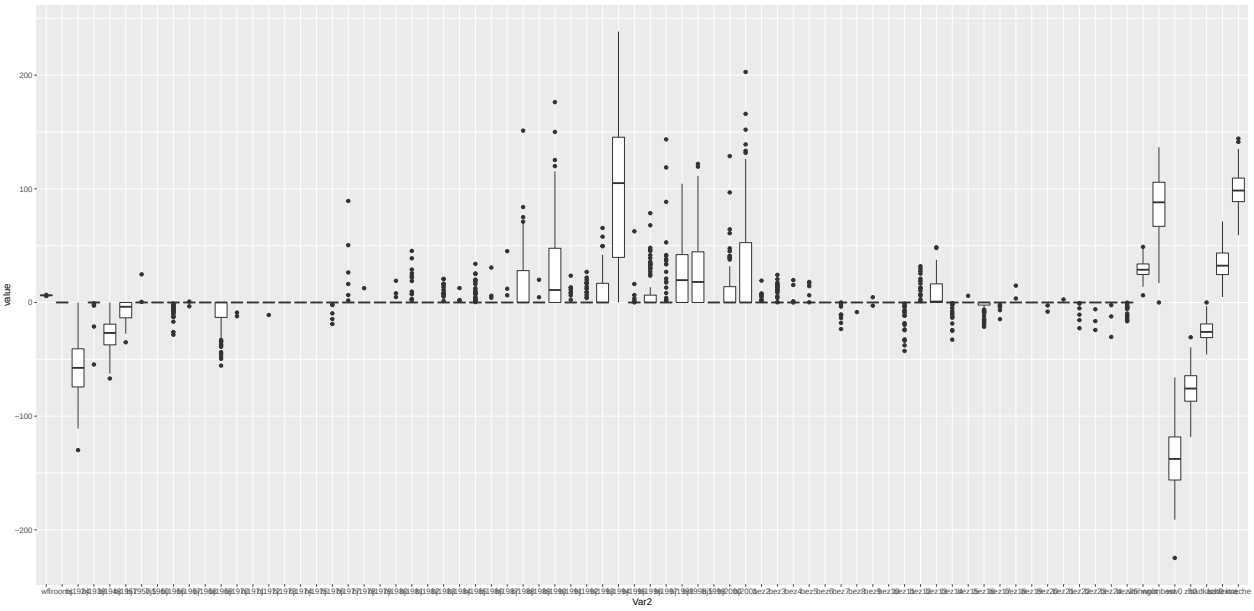
vanilla and ridge regression were not capable to give explainable outcomes. Likewise, the result of the lasso was contra-intuitive in the unclear handling of the factorial varibles. Finally, the group lasso where a factorial varible can be arranged together leads to an interpretable shrinkage conclusion, where all variables except the factorial are selected for the optimal hyperparemeter choice.

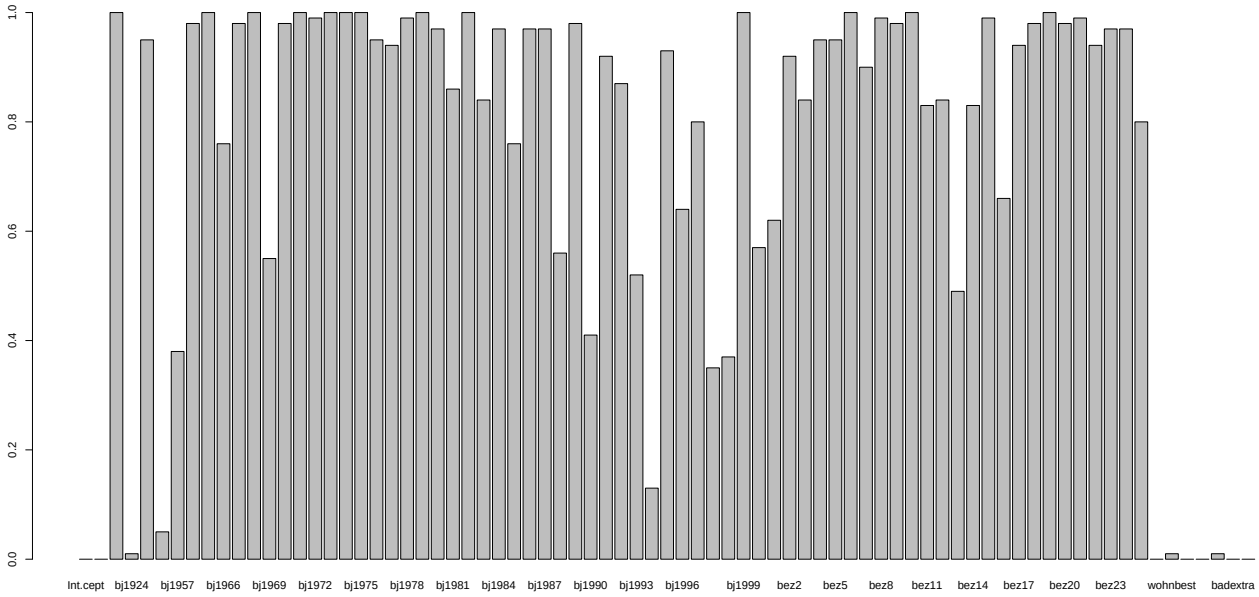
Inference

The data set is too small to divide it into a training and a test set.

Bootstrap can be applied here to find the proportion of times each element in the coefficients vector of being zero. So it is a way of validation. By looking at the boxplot and barplot which show how many percentages of that variable of being zero, one can tell that the majority of the coefficients for those year covariates are zero while the rest of the dominating factors are non zero which reflects their significance as well.

So to conclude, one can tell group lasso seems working suitably for this specific case where its categorical covariates have many groups. Ridge regression does not shrink any parameters dramatically while lasso does a bit on the shrinkage, but the most shrinked contributation is from group lasso.




0.0
0.2
0.4
0.6
0.8
1.0
Int.cept
bj1924
bj1957
bj1966
bj1969
bj1972
bj1975
bj1978
bj1981
bj1984
bj1987
bj1990
bj1993
bj1996
bj1999
bez2
bez5
bez8
bez11
bez14
bez17
bez20
bez23
wohnbest
badextra