Prof. Dr. Markus Bibinger
AG Quantitatives Risiko und
hochdimensionale Finanzdaten
Wintersemester 2024/25
Julius-Maximilians-Universität Würzburg

Project regularized regression: Credit data

(a) Install the packages glmnet and ISLR2 and consider the credit data set str(Credit). The response is balance (average credit card debt for each individual). Asses first heuristically the relation between each explanatory variable and the response separately.

(b) Consider a regression with $p = 3$ and the numeric explanatory variables income, limit and rating. Standardize these three variables with their empirical standard deviations and then combine them in a design matrix. Compute ridge regression estimates for this regression based on glmnet (..., alpha = 0, intercept =T,lambda =...) across different penalty parameters

```
grid <- rev(10^seq(4, -2, length = 10000)).
```

Illustrate the estimates jointly in one plot. Compute the lasso estimates for the same regression with the same grid of penalty parameters. Create an analogous plot.
Compare the ratios of the L_2-norms of the ridge regression and the least squares estimates across the grid of penalty parameters. Perform the comparison first with and then without the intercept parameter. Illustrate the results in plots and discuss the results.
Hint: The command coef( ) can be useful.

(c) One could reasonably have measured income in dollars instead of thousands of dollars, which would result in a rescaling of the observed values of income by a factor of 1,000. Consider explanatory variables income, limit and rating, here without standardization. How should rescaling influence least squares and the regularized regression results theoretically? Compute the estimates before and after rescaling, the lasso and ridge regression once more with glmnet and with $\lambda = 100$, and interpret the outcomes.

(d) Perform the regressions with the complete designs given in the data. Use the command predict ( ) to compare predictions based on the three different regression estimates, least squares, ridge regression and lasso. Make predictions for the explanatory variables from the 400th individual in the sample as an input. Use data driven choices of λ here with

```
cv.glmnet( )$lambda.min.
```

Hint: The command model.matrix( ) can be useful.

(e) Select λ for the lasso within the grid (0:100)/10 by splitting the data randomly in one half of training and another half of test sample, the random decompositions iterated $M = 1{,}000$ times, and minimizing the overall empirical L_2-prediction error of the forecasts for the test samples based on the regressions with the training samples. Illustrate the dependence of the empirical L_2-prediction error on λ in a plot. Compare to analogous empirical L_2-prediction errors based on least squares estimates and of simply fitting a model with just an intercept.

The project results are discussed in a presentation and a concise report is submitted.