Regularized regression

Where regression analysis meets machine learning

Who am I?

Rune Hjorth Nielsen

- Data Scientist and AI Specialist
  - Nordic CxP team for AI and Advanced analytics
- Based in Aarhus, Denmark
- PhD in Mathematical Economics
  - Applied time series modelling estimated using regularized regression
- RuneHjorth.Nielsen@sas.com
- www.linkedin.com/in/runehnielsen/

Magic in advanced analytics?

Agenda

- Some versions of regularized regression
- Parameter tuning
- One way of estimating in SAS
- (Extreme) Use cases

Different regularization approaches

Objective functions

Ridge regression

Hoerl and Kennard (1970)

$$\hat{\beta}^{Ridge} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - x_i{}^T \beta)^2 + \lambda \sum_{j=1}^{p} \beta_j{}^2$$

Main use case:

- Fear of overfitting

Lasso regression

Tibshirani (1996)

$$\hat{\beta}^{Lasso} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - x_i{}^T \beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

Additional use case:

- Assumed sparsity in models

Adaptive lasso

Zou (2006)

$$\hat{\beta}^{Ada\_Lasso} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - x_i{}^T \beta)^2 + \lambda \sum_{j=1}^{p} \widehat{w}_j |\beta_j|$$

$$= \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - x_i{}^T \beta)^2 + \lambda \sum_{j=1}^{p} \frac{1}{\hat{\beta}_j^{Lasso}} |\beta_j|$$

Additional use case:

- Oracle properties

Elasticnet regression

Zou & Hastie (2005)

$$\hat{\beta}^{Elasticnet} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - x_i{}^T \beta)^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j{}^2$$

Additional use case:

- Handling high-correlated data

Estimating the lambda

Estimating Lambda

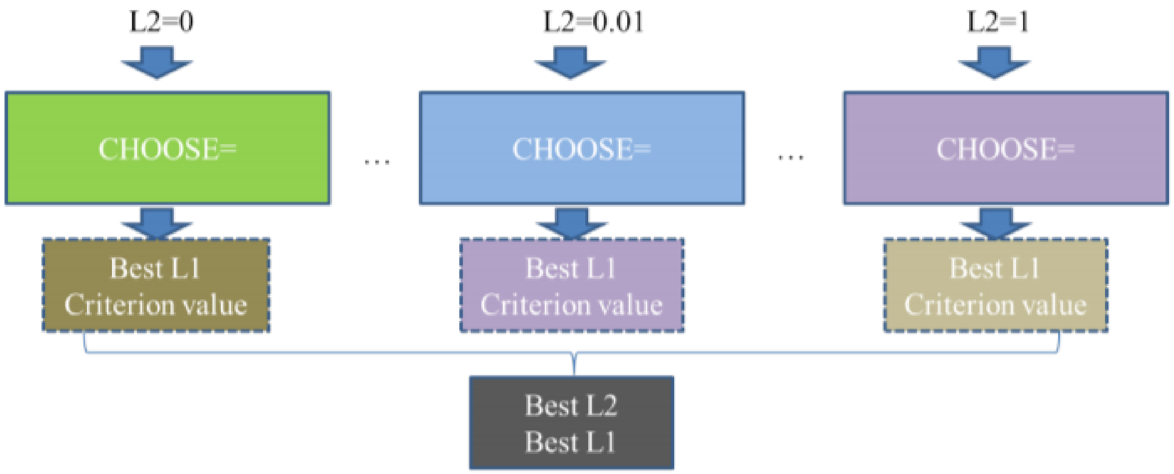
Test a sequence of different lambdas. Choose on some model performance.

- Model statistics (e.g. predicted residual error sum of squares)
  - Often done through cross validation
- Information criteria (e.g BIC)
  - Might be better at handling a specific structure in the data

Estimating Lambda

1. Define steps of λ_2
2. Decide on a λ_1 based on chosen criterion
3. Choose best combination of λ_1 and λ_2 base on the chosen criterion

Figure 51.12 Estimation of the Ridge Regression Parameter λ_2 (L2) in the Elastic Net Method



Why use regularized regression?

Pros

- Oracle properties
  - Variable selection
- Reducing overfitting
- Estimating more parameters than sample points
  - Highly relevant for small samples

Cons

- Bias!
- Parameter tuning
  - Grows with complexity
- Computationally expensive
  - No analytical solution

Lasso estimation example

Lasso estimation example

The simulated data

```
%let N = 100; /* size of each sample */
%let ZVar = 1;
data Reg2(keep=y x0 x1 x2 x3 x4 a1 a2 b1 b2 c1 c2);
call streaminit(123);
do i = 1 to &N;
    x0 = rand("Normal",5,5);
    x1 = rand("Normal",5,5);
    x2 = rand("Normal",5,5);
    x3 = rand("Normal",5,5);
    x4 = rand("Normal",5,5);
    a1 = rand("Normal",5,5);
    a2 = a1 + rand("normal",0,&ZVar);
    b1 = rand("Normal",5,5);
    b2 = b1 + rand("normal",0,&ZVar);
    c1 = rand("Normal",5,5);
    c2 = c1 + rand("normal",0,&ZVar);
    eps = rand("Normal", 0, 10);
    y = 1 + 0*x0 + 1*x1 + 2*x2 + 3*x3 + 0*x4 + 1*a1 + 2*a2 + 1*b1 + 2*b2 + 1*c1 + 2*c2 + eps;
    output;
end;
run;
```

Lasso estimation example

Simulated model

Expected model:

$$y = 1 + 1x_1 + 2x_2 + 3x_3 + 1a_1 + 2a_2 + 1b_1 + 2b_2 + 1c_1 + 2c_2 + \varepsilon$$

Lasso estimation example

Base SAS implementation

```
proc glmselect data = Reg2 plots = all;
model y = x0 x1 x2 x3 x4 a1 a2 b1 b2 c1 c2 /
         selection=lasso (choose=cv stop=none) cvmethod=random(10);
run;
```

Lasso estimation example

Model selection

Data Set	WORK.REG2
Dependent Variable	y
Selection Method	LASSO
Stop Criterion	None
Choose Criterion	Cross Validation
Cross Validation Method	Random
Cross Validation Fold	10
Effect Hierarchy Enforced	None
Random Number Seed	751790182

Number of Observations Read	100
Number of Observations Used	100

Dimensions	
Number of Effects	12
Number of Parameters	12

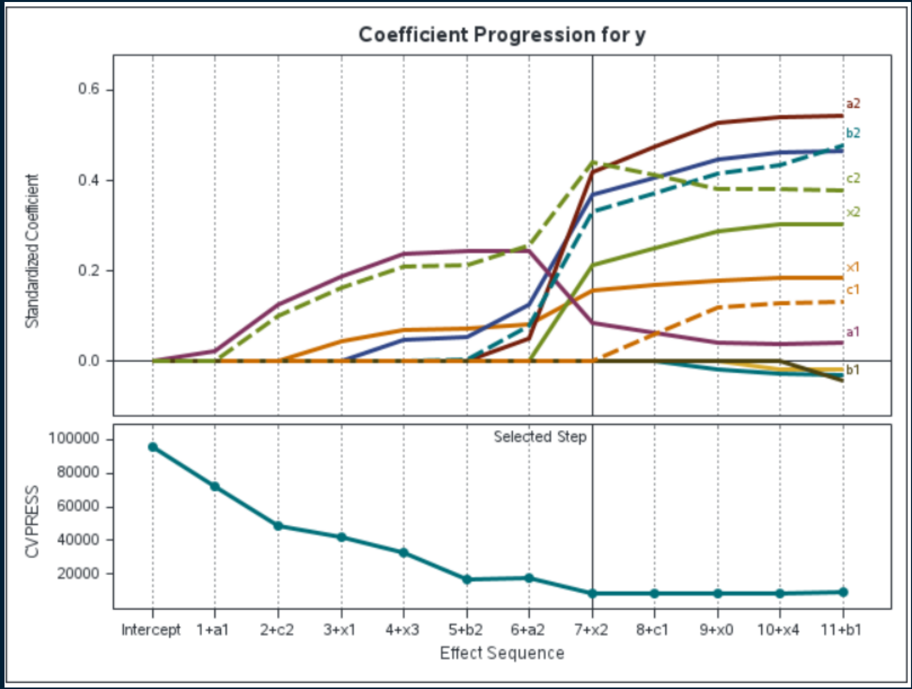
LASSO Selection Summary

Step	Effect Entered	Effect Removed	Number Effects In	CV PRESS
0	Intercept		1	95934.9734
1	a1		2	72567.3798
2	c2		3	49023.3523
3	x1		4	42304.6625
4	x3		5	33111.4990
5	b2		6	17155.5245
6	a2		7	17333.2071
7	x2		8	8282.4601*
8	c1		9	8306.7816
9	x0		10	8396.2930
10	x4		11	8742.3712
11	b1		12	9060.4536

* Optimal Value of Criterion

Lasso estimation example

Model selection



Lasso estimation example

Estimated parameters

Root MSE	10.33696
Dependent Mean	81.87112
R-Square	0.8963
Adj R-Sq	0.8884
AIC	576.80697
AICC	578.80697
SBC	495.64833
CV PRESS	8282.46011

Parameter Estimates		
Parameter	DF	Estimate
Intercept	1	17.251625
x1	1	1.080064
x2	1	1.247643
x3	1	2.273183
a1	1	0.460773
a2	1	2.245478
b2	1	2.105657
c2	1	2.529493

Lasso estimation example

Using different tuning approach

```
proc glmselect data = Reg2 plots = all;
model y = x0 x1 x2 x3 x4 a1 a2 b1 b2 c1 c2 /
          selection=lasso (choose=BIC stop=none);
run;
```

Lasso estimation example

Compare cross validation vs BIC

LASSO Selection Summary

Step	Effect Entered	Effect Removed	Number Effects In	CV PRESS
0	Intercept		1	95934.9734
1	a1		2	72567.3798
2	c2		3	49023.3523
3	x1		4	42304.6625
4	x3		5	33111.4990
5	b2		6	17155.5245
6	a2		7	17333.2071
7	x2		8	8282.4601*
8	c1		9	8306.7816
9	x0		10	8396.2930
10	x4		11	8742.3712
11	b1		12	9060.4536

* Optimal Value of Criterion

LASSO Selection Summary

Step	Effect Entered	Effect Removed	Number Effects In	BIC
0	Intercept		1	685.8971
1	a1		2	683.7400
2	c2		3	664.1220
3	x1		4	648.1638
4	x3		5	632.5531
5	b2		6	631.0765
6	a2		7	607.1485
7	x2		8	473.5439
8	c1		9	457.7546
9	x0		10	449.8742*
10	x4		11	450.8733
11	b1		12	453.0704

* Optimal Value of Criterion

(Extreme) use cases

Extreme use cases

Uniejewski, Bartosz & Marcjasz, Grzegorz & Weron, Rafał. (2019)

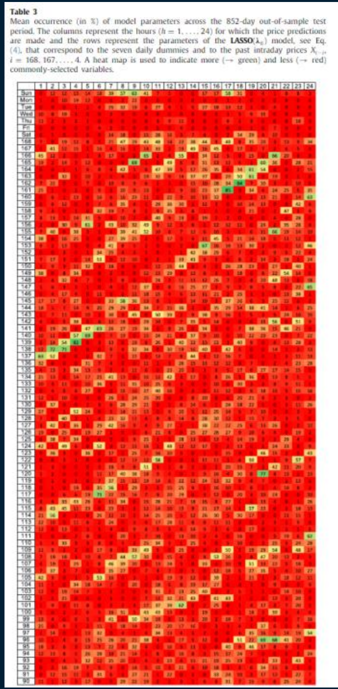


Table 3
Mean occurrence (in %) of model parameters across the 852-day out-of-sample test period. The columns represent the hours ($h = 1, \ldots, 24$) for which the price predictions are made and the rows represent the parameters of the **LASSO**(λ_8) model, see Eq. (4), that correspond to the seven daily dummies and to the past intraday prices X_{t-i}, $i = 168, 167, \ldots, 4$. A heat map is used to indicate more (→ green) and less (→ red) commonly-selected variables.

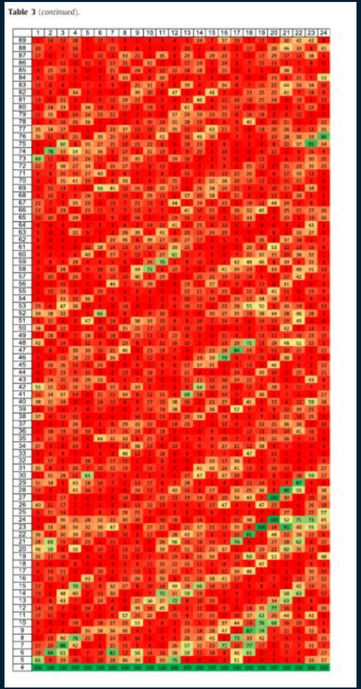


Table 3 (*continued*).

Extreme use cases

Less observations than parameters

```
%let N = 5; /* size of each sample */
%let ZVar = 1;
data Reg2(keep=y x0 x1 x2 x3 x4 a1 a2 b1 b2 c1 c2);
call streaminit(123);
do i = 1 to &N;
    x0 = rand("Normal",5,5);
    x1 = rand("Normal",5,5);
    x2 = rand("Normal",5,5);
    x3 = rand("Normal",5,5);
    x4 = rand("Normal",5,5);
    a1 = rand("Normal",5,5);
    a2 = a1 + rand("normal",0,&ZVar);
    b1 = rand("Normal",5,5);
    b2 = b1 + rand("normal",0,&ZVar);
    c1 = rand("Normal",5,5);
    c2 = c1 + rand("normal",0,&ZVar);
    eps = rand("Normal", 0, 10);
    y = 1 + 0*x0 + 1*x1 + 2*x2 + 3*x3 + 0*x4 + 1*a1 + 2*a2 + 1*b1 + 2*b2 + 1*c1 + 2*c2 + eps;
    output;
end;
run;
```

Extreme use cases

Less observations than parameters

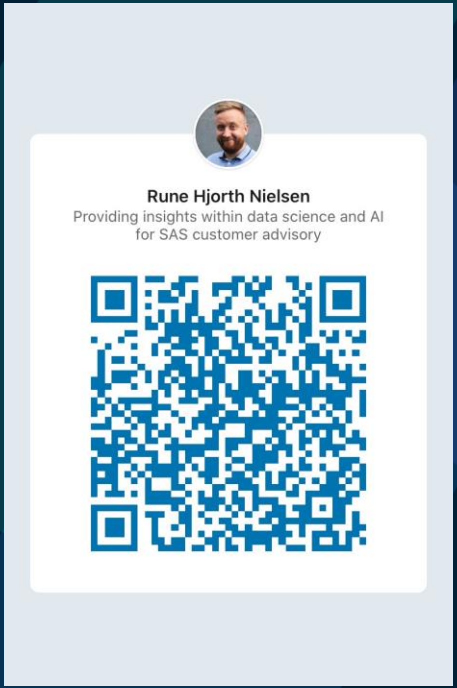
LASSO Selection Summary				
Step	Effect Entered	Effect Removed	Number Effects In	CV PRESS
0	Intercept		1	4531.4006
1	x1		2	1439.4592
2	a2		3	80.0936*
3	b1		4	414.1813
4	c2		5	414.1813

* Optimal Value of Criterion

References

- Photo by Aron Visuals on Unsplash
- Uniejewski, Bartosz & Marcjasz, Grzegorz & Weron, Rafał. (2019). Understanding intraday electricity markets: Variable selection and very short-term price forecasting using LASSO. International Journal of Forecasting. 35. 10.1016/j.ijforecast.2019.02.001.

Questions?


Rune Hjorth Nielsen
Providing insights within data science and AI for SAS customer advisory