Machine Learning for econometrics

Flexible models for tabular data

Matthieu Doutreligne

A lot of today's content is taken from the excellent sklearn mooc (Estève et al., 2022)

Reminder from previous session

- Statistical learning 101: bias-variance trade-off
- Regularization for linear models: Lasso, Ridge, Elastic Net
- Transformation of variables: polynomial regression

Reminder from previous session

- Statistical learning 101: bias-variance trade-off
- Regularization for linear models: Lasso, Ridge, Elastic Net
- Transformation of variables: polynomial regression

🤔 But… How to select the best model? the best hyper-parameters?

Contents

Contents

Example with the Wage dataset

- Raw dataset: (N=534, p=11)

EDUCATION	SOUTH	SEX	EXPERIENCE	UNION	WAGE	AGE	RACE	OCCUPATION	SECTOR	MARR
8	no	female	21	not_member	5.10	35	Hispanic	Other	Manufacturing	Married
9	no	female	42	not_member	4.95	57	White	Other	Manufacturing	Married
12	no	male	1	not_member	6.67	19	White	Other	Manufacturing	Unmarried
12	no	male	4	not_member	4.00	22	White	Other	Other	Unmarried
12	no	male	17	not_member	7.50	35	White	Other	Other	Married

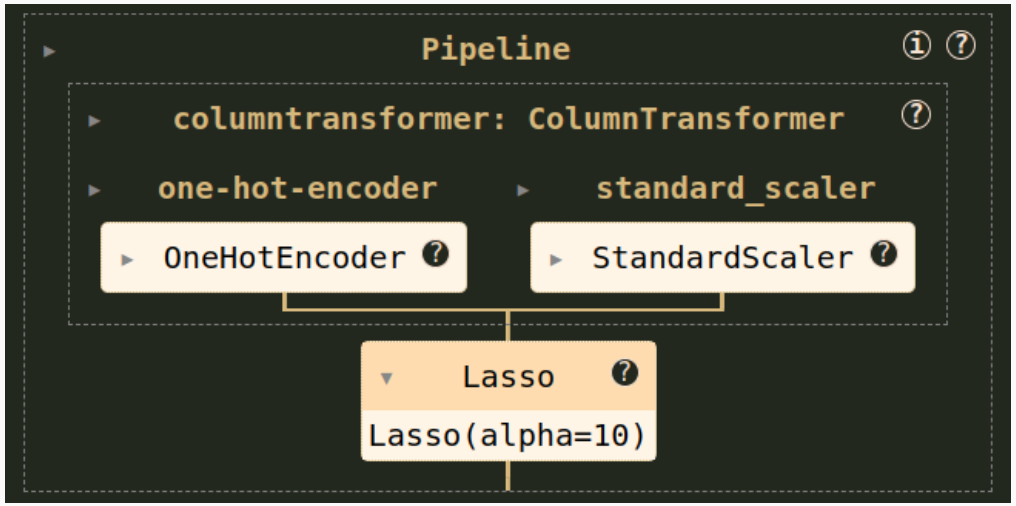
Example with the Wage dataset

- Raw dataset: (N=534, p=11)
- Transformation: encoding categorical data, scaling numerical data: (N=534, p=23)

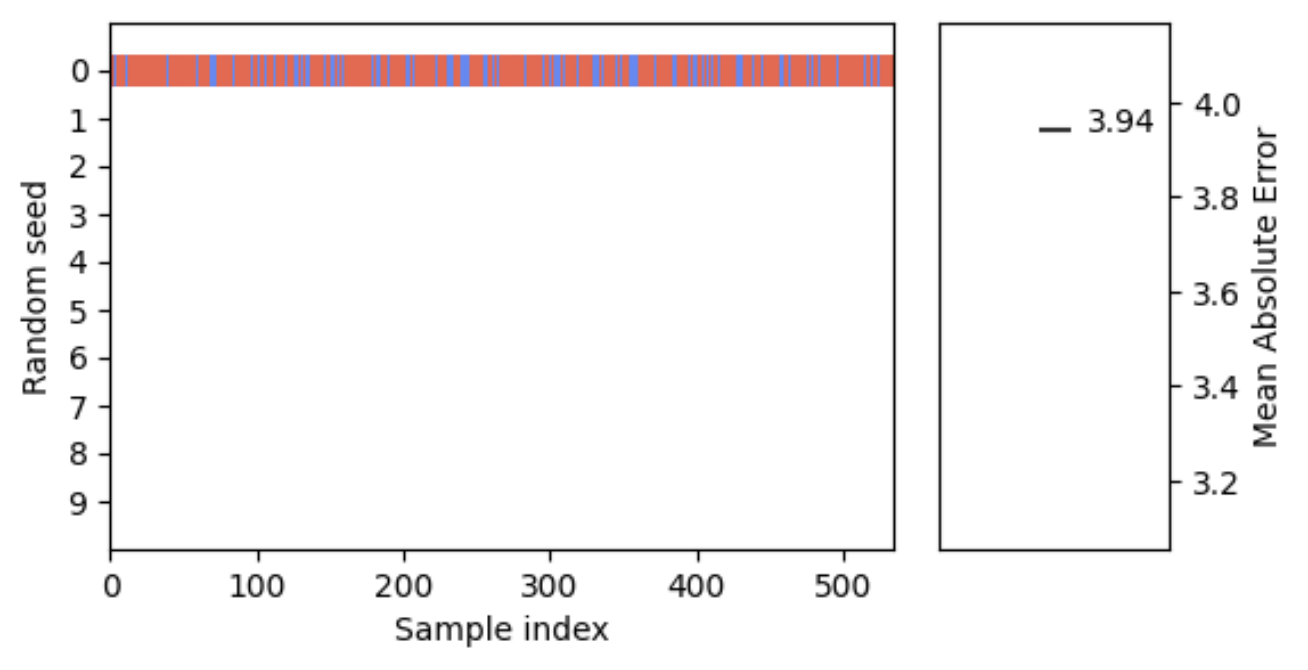
one-hot-encoder__SOUTH_no	one-hot-encoder__SOUTH_yes	one-hot-encoder__SEX_female	one-hot-encoder__SEX_male	one-hot-encoder__UNION_member	encoder__UNION_not
1.0	0.0	1.0	0.0	0.0	
1.0	0.0	1.0	0.0	0.0	
1.0	0.0	0.0	1.0	0.0	
1.0	0.0	0.0	1.0	0.0	
1.0	0.0	0.0	1.0	0.0	

Example with the Wage dataset

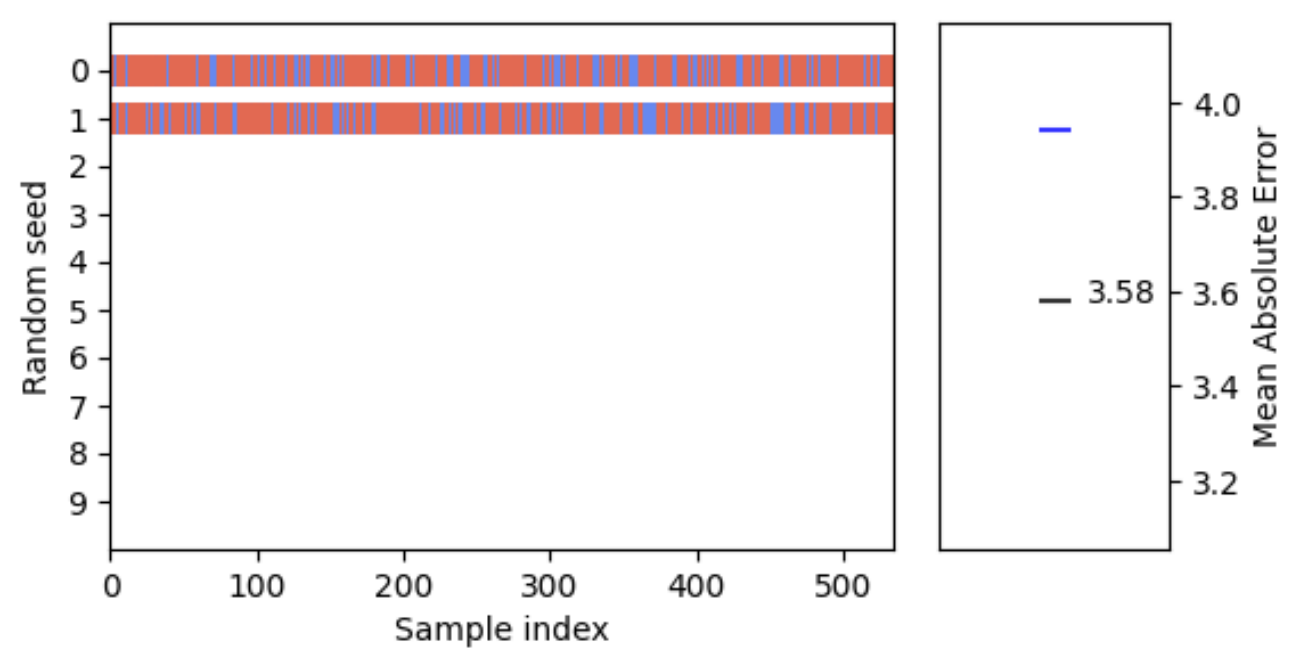
- Raw dataset: (N=534, p=11)
- Transformation: encoding categorical data, scaling numerical data: (N=534, p=23)
- Regressor: Lasso with regularization parameter ($\alpha = 10$)



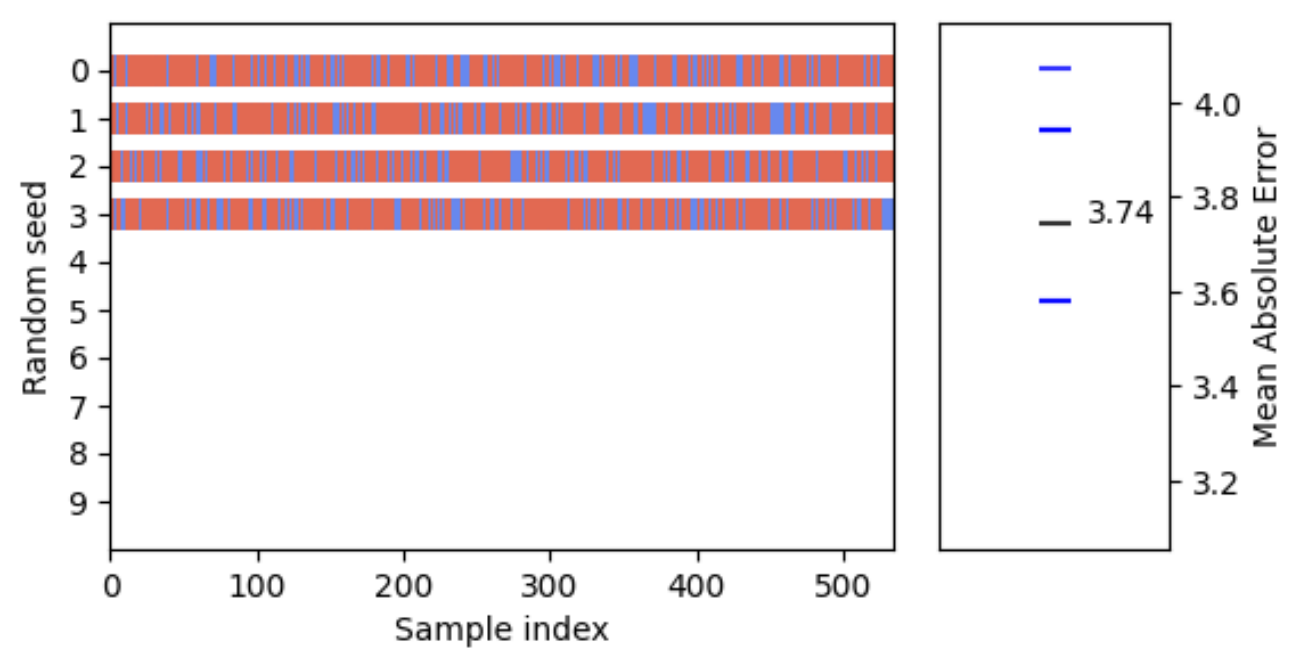
Splitting once: In red, the training set, in blue, the test set



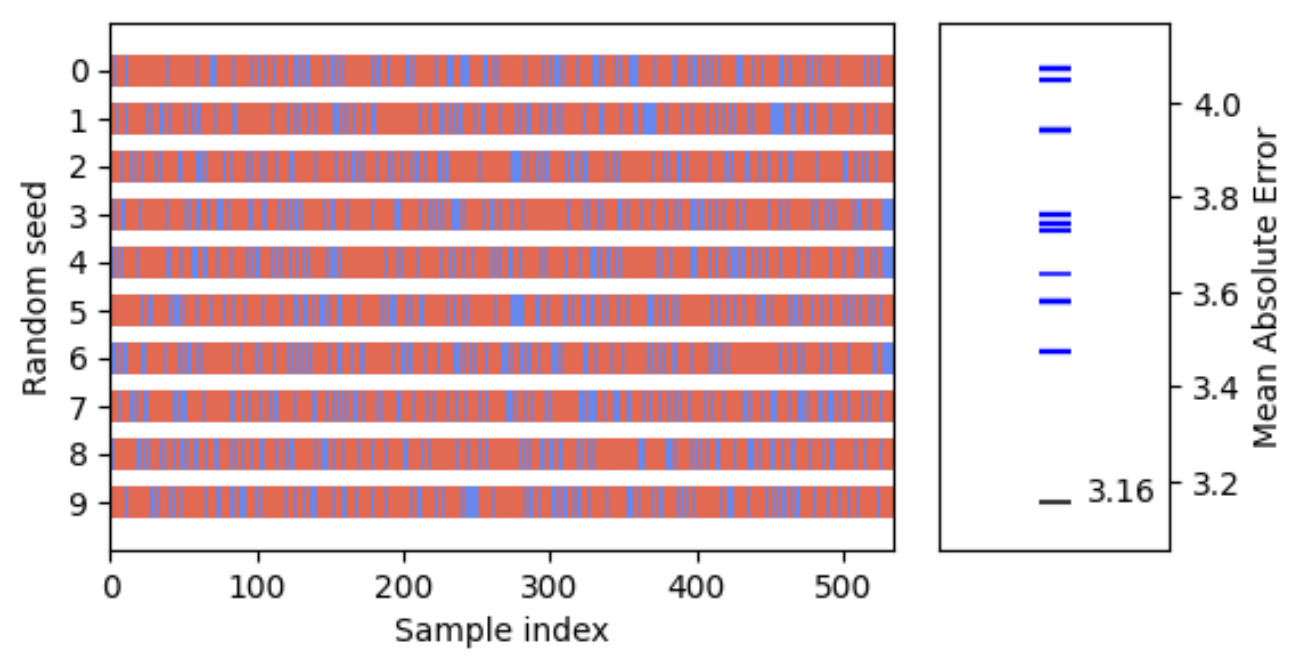
But we could have chosen another split ! Yielding a different MAE



And another split...



Splitting ten times



🎉 **Distribution of MAE:** 3.71 ± 0.26

Repeated exclusive train/test splits = Cross-validation

Practical usage with sklearn: `cross_validate`.

```python
from sklearn.model_selection import cross_validate
cv_results = cross_validate(
    regressor, data, target, cv=5,
    scoring="neg_mean_absolute_error"
)
```

Repeated exclusive train/test splits = Cross-validation

Practical usage with sklearn: `cross_validate`.

```python
from sklearn.model_selection import cross_validate
cv_results = cross_validate(
    regressor, data, target, cv=5,
    scoring="neg_mean_absolute_error"
)
```

- 🙂 Robustly estimate generalization performance.
- 🤩 Estimate data variability of the performance : bigger source of variation (Bouthillier et al., 2021).
- 🚀 Let's use it to select the best models among several candidates!

Contents

Python hands-on

To your notebooks 👨‍💻!

- url: https://github.com/strayMat/causal-ml-course/tree/main/notebooks

Bibliography

Bouthillier, X., Delaunay, P., Bronzi, M., Trofimov, A., Nichyporuk, B., Szeto, J., Mohammadi Sepahvand, N., Raff, E., Madan, K., Voleti, V., & others. (2021). Accounting for variance in machine learning benchmarks. Proceedings of Machine Learning and Systems, 3, *747–769.*

Estève, L., Lemaitre, G., Grisel, O., Varoquaux, G., Amor, A., Lilian, Rospars, B., Schmitt, T., Liu, L., Kinoshita, B. P., hackmd-deploy, ph4ge, Steinbach, P., Boucaud, A., Muite, B., Boisberranger, J. du, Notter, M., Pierre, P, S., … parmentelat. (2022). INRIA/scikit-learn-mooc: Third MOOC session. Zenodo. *https://doi.org/10.5281/zenodo.7220307*