Machine Learning for Official Statistics and SDGs

Third Live Lecture (Webinar):

Starts in **15** minutes

Christophe Bontemps, UN SIAP

UNITED NATIONS SIAP
Statistical Institute for Asia and the Pacific

Machine Learning for Official Statistics and SDGs

Third Live Lecture (Webinar):

Starts in **10** minutes

Christophe Bontemps, UN SIAP

UNITED NATIONS SIAP Statistical Institute for Asia and the Pacific

Machine Learning for Official Statistics and SDGs

Third Live Lecture (Webinar): Starts in **5** minutes

Christophe Bontemps, UN SIAP

UNITED NATIONS
SIAP
Statistical Institute for Asia and the Pacific

Machine Learning for Official Statistics and SDGs

On Regression

UNITED NATIONS
SIAP
Statistical Institute for Asia and the Pacific

[- Reminder -]

- Mute yourself always!

[- Reminder -]

- Mute yourself **always**!
- The lecture is recorded

[- REMINDER -]

- Mute yourself always!
- The lecture is recorded
- Ask questions in the chat

[- Agenda -]

- Introduction

[- AGENDA -]

- Introduction
- Regression (*in a Machine learning Framework*)

[- Agenda -]

- Introduction
- Regression (*in a Machine learning Framework*)
- Q&A

[- Agenda -]

- Introduction
- Regression (*in a Machine learning Framework*)
- Q&A
- Next week

[Linear Regression]

Multivariate Linear Regression is one of the most popular tool

[Linear Regression]

Multivariate Linear Regression is one of the most popular tool

- Can be used with both continuous or discrete variables (categories)

[Linear Regression]

Multivariate Linear Regression is one of the most popular tool

- Can be used with both continuous or discrete variables (categories)
- Has to be well defined, need to verify some hypothesis

[LINEAR REGRESSION]

Multivariate Linear Regression is one of the most popular tool

- Can be used with both continuous or discrete variables (categories)
- Has to be well defined, need to verify some hypothesis

  Expressed as:

$$y = \beta_0 + x'\beta + \varepsilon \qquad E(\varepsilon|x) = 0$$

[Linear Regression]

Multivariate Linear Regression is one of the most popular tool

- Can be used with both continuous or discrete variables (categories)
- Has to be well defined, need to verify some hypothesis

  Expressed as:

$$y = \beta_0 + x'\beta + \varepsilon \qquad E(\varepsilon|x) = 0$$

  with possibly many regressors x_j

[Linear Regression]

Multivariate Linear Regression is one of the most popular tool

- Can be used with both continuous or discrete variables (categories)
- Has to be well defined, need to verify some hypothesis

  Expressed as:

  $$y = \beta_0 + x'\beta + \varepsilon \qquad E(\varepsilon|x) = 0$$

  with possibly many regressors x_j
- βs are solutions of:

$$\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \beta_0 - x_i'\beta\right)^2$$

[Linear Regression: Centering variables]

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \varepsilon \qquad E(\varepsilon|x) = 0$$

[Linear Regression: Centering variables]

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \varepsilon \qquad E(\varepsilon|x) = 0$$

- β_j is the "*ceteris paribus*" marginal effect of x_j on y.

[Linear Regression: Centering variables]

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \varepsilon \qquad E(\varepsilon | x) = 0$$

- β_j is the "*ceteris paribus*" marginal effect of x_j on y.

↪ *when x_j increases by one unit, then y increase by β_j units.*

[Linear Regression: Centering variables]

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \varepsilon \qquad E(\varepsilon | x) = 0$$

- β_j is the "*ceteris paribus*" marginal effect of x_j on y.

↪ *when x_j increases by one unit, then y increase by β_j units.*

- β_0 is the mean of y **if all x_j are equal to zero**

$$\beta_0 = E(y) - \beta_1 E(x_1) - \ldots - \beta_k E(x_k)$$

[Linear Regression: Centering variables]

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \varepsilon \qquad E(\varepsilon|x) = 0$$

- β_j is the "*ceteris paribus*" marginal effect of x_j on y.

↪ *when x_j increases by one unit, then y increase by β_j units.*

- β_0 is the mean of y **if all x_j are equal to zero**

$$\beta_0 = E(y) - \beta_1 E(x_1) - \ldots - \beta_k E(x_k)$$

- Centering the variables has no effect on the coefficients

$$y = \alpha_0 + \beta_1(x_1 - E(x_1)) + \ldots + \beta_k(x_k - E(x_k)) + \varepsilon \qquad E(\varepsilon|x) = 0$$

[Linear Regression: Centering variables]

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \varepsilon \qquad E(\varepsilon|x) = 0$$

- β_j is the "*ceteris paribus*" marginal effect of x_j on y.

$\hookrightarrow$ *when x_j increases by one unit, then y increase by β_j units.*

- β_0 is the mean of y **if all x_j are equal to zero**

$$\beta_0 = E(y) - \beta_1 E(x_1) - \ldots - \beta_k E(x_k)$$

- Centering the variables has no effect on the coefficients

$$y = \alpha_0 + \beta_1(x_1 - E(x_1)) + \ldots + \beta_k(x_k - E(x_k)) + \varepsilon \qquad E(\varepsilon|x) = 0$$

Except: α_0 is the mean of y **if all x_j are equal their mean**

[Linear Regression: Scaling]

We can also scale each variable by its own standard deviation to obtain

$$y = \alpha_0 + \gamma_1 \tilde{x}_1 + \ldots + \gamma_k \tilde{x}_k + \varepsilon \qquad E(\varepsilon|x) = 0$$

[Linear Regression: Scaling]

We can also scale each variable by its own standard deviation to obtain

$$y = \alpha_0 + \gamma_1 \tilde{x}_1 + \ldots + \gamma_k \tilde{x}_k + \varepsilon \qquad E(\varepsilon|x) = 0$$

where

$$\tilde{x} = \frac{x - E(x)}{\sigma_x}$$

[Linear Regression: Scaling]

We can also scale each variable by its own standard deviation to obtain

$$y = \alpha_0 + \gamma_1 \tilde{x}_1 + \ldots + \gamma_k \tilde{x}_k + \varepsilon \qquad E(\varepsilon | x) = 0$$

where

$$\tilde{x} = \frac{x - E(x)}{\sigma_x}$$

- Now γ_j is the *ceteris paribus* marginal effect of $\tilde{x}_j$ on y.

[Linear Regression: Scaling]

We can also scale each variable by its own standard deviation to obtain

$$y = \alpha_0 + \gamma_1 \tilde{x}_1 + \ldots + \gamma_k \tilde{x}_k + \varepsilon \qquad E(\varepsilon | x) = 0$$

where

$$\tilde{x} = \frac{x - E(x)}{\sigma_x}$$

- Now γ_j is the *ceteris paribus* marginal effect of $\tilde{x}_j$ on y.

$\hookrightarrow$ *when $\tilde{x}_j$ increases by one standard deviation, y increases by γ_j units*

[Linear Regression: Scaling]

We can also scale each variable by its own standard deviation to obtain

$$y = \alpha_0 + \gamma_1 \tilde{x}_1 + \ldots + \gamma_k \tilde{x}_k + \varepsilon \qquad E(\varepsilon | x) = 0$$

where

$$\tilde{x} = \frac{x - E(x)}{\sigma_x}$$

- Now γ_j is the *ceteris paribus* marginal effect of $\tilde{x}_j$ on y.

↪ *when $\tilde{x}_j$ increases by one standard deviation, y increases by γ_j units*

- The goal is to have variables and coefficients that are comparable

[Linear Regression: Example]

[Linear Regression: Example]

- Example on a regression model with few variables

	Est.	S.E.	t val.	p
(Intercept)	4.390	0.136	32.204	0.000
Trans	-0.002	0.001	-1.603	0.110
HighTrans	0.013	0.003	3.945	0.000
Checks	0.008	0.005	1.534	0.126
Years	0.096	0.008	11.579	0.000

Initial regression with original values

[Linear Regression: Example]

- Example on a regression model with few variables

	Est.	S.E.	t val.	p
(Intercept)	5.927	0.039	150.243	0.000
Trans	-0.002	0.001	-1.603	0.110
HighTrans	0.013	0.003	3.945	0.000
Checks	0.008	0.005	1.534	0.126
Years	0.096	0.008	11.579	0.000

Regression with centred variables

[Linear Regression: Example]

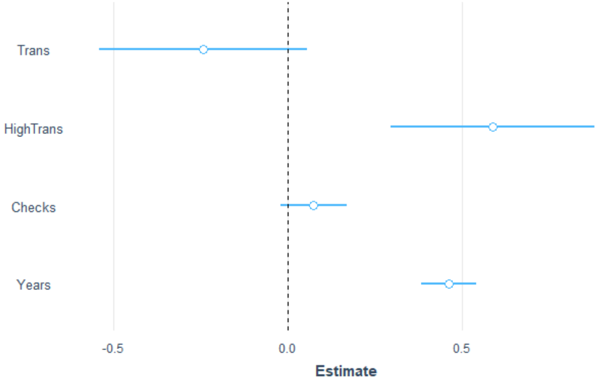
- Example on a regression model with few variables

	Est.	S.E.	t val.	p
(Intercept)	5.927	0.039	150.243	0.000
Trans	-0.243	0.152	-1.603	0.110
HighTrans	0.586	0.149	3.945	0.000
Checks	0.074	0.048	1.534	0.126
Years	0.462	0.040	11.579	0.000

Regression with scaled variables

[Linear Regression: Example]

- Example on a regression model with few variables
  The goal is to have *comparable* effects (same range)



Visual regression with scaled variables

[Problems in Linear Regression]

Collinearity of regressors is a big issue in regression

[Problems in Linear Regression]

Collinearity of regressors is a big issue in regression

- Redundant predictors add more complexity than information

[Problems in Linear Regression]

Collinearity of regressors is a big issue in regression

- Redundant predictors add more complexity than information
- Highly correlated predictors result in unstable estimation

[Problems in Linear Regression]

Collinearity of regressors is a big issue in regression

- Redundant predictors add more complexity than information
- Highly correlated predictors result in unstable estimation
- Highly correlated predictors worsen predictability

[Problems in Linear Regression]

Collinearity of regressors is a big issue in regression

- Redundant predictors add more complexity than information
- Highly correlated predictors result in unstable estimation
- Highly correlated predictors worsen predictability

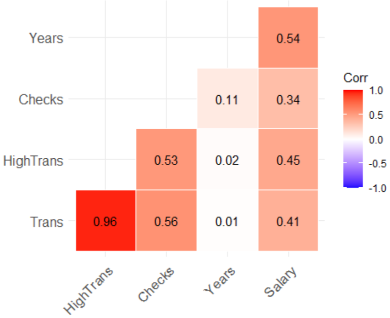
$\hookrightarrow$ Correlation plot

[Problems in Linear Regression]

Collinearity of regressors is a big issue in regression

- Redundant predictors add more complexity than information
- Highly correlated predictors result in unstable estimation
- Highly correlated predictors worsen predictability

$\hookrightarrow$ Correlation plot



[Problems in Linear Regression]

Which variable should be removed?

[Problems in Linear Regression]

Which variable should be removed?

Variance Inflation Factor

[Problems in Linear Regression]

Which variable should be removed?

Variance Inflation Factor

- Measure of **multi-collinearity** between variables

[Problems in Linear Regression]

Which variable should be removed?

Variance Inflation Factor

- Measure of **multi-collinearity** between variables
- VIF for x_j is calculated by running a regression of x_j on all other regressors, computing the R_j^2 and use the formula:

[Problems in Linear Regression]

Which variable should be removed?

Variance Inflation Factor

- Measure of **multi-collinearity** between variables
- VIF for x_j is calculated by running a regression of x_j on all other regressors, computing the R_j^2 and use the formula:

$$VIF_j = \frac{1}{1 - R_j^2}$$

[Problems in Linear Regression]

Which variable should be removed?

Variance Inflation Factor

- Measure of **multi-collinearity** between variables
- VIF for x_j is calculated by running a regression of x_j on all other regressors, computing the R_j^2 and use the formula:

$$VIF_j = \frac{1}{1 - R_j^2}$$

$\hookrightarrow$ A $VIF_j = 1$ indicates no collinearity

[Problems in Linear Regression]

Which variable should be removed?

Variance Inflation Factor

- Measure of **multi-collinearity** between variables
- VIF for x_j is calculated by running a regression of x_j on all other regressors, computing the R_j^2 and use the formula:

$$VIF_j = \frac{1}{1 - R_j^2}$$

↪ A $VIF_j = 1$ indicates no collinearity

↪ A $VIF \geq 10$ is considered as large and problematic

[Solutions to multi-collinearity]

2 solutions to multi-colinearity:

[Solutions to multi-collinearity]

2 solutions to multi-colinearity:

- Create new variables from the ones that are collinear

[Solutions to multi-collinearity]

2 solutions to multi-colinearity:

- Create new variables from the ones that are collinear

$\hookrightarrow$ using *e.g.* Principal Components Analysis

[Solutions to multi-collinearity]

2 solutions to multi-colinearity:

- Create new variables from the ones that are collinear

↪ using *e.g.* Principal Components Analysis

- Remove some variables

[Using *VIF* to remove collinearity]

[Using *VIF* to remove collinearity]

Computing VIFs for all x_js

[Using *VIF* to remove collinearity]

Computing VIFs for all x_js

	Est.	S.E.	t val.	p	VIF
(Intercept)	5.93	0.04	150.24	0.00	NA
Trans	-0.24	0.15	-1.60	0.11	14.71
HighTrans	0.59	0.15	3.94	0.00	14.15
Checks	0.07	0.05	1.53	0.13	1.47
Years	0.46	0.04	11.58	0.00	1.02

[USING *VIF* TO REMOVE COLLINEARITY]

Computing VIFs for all x_js

	Est.	S.E.	t val.	p	VIF
(Intercept)	5.93	0.04	150.24	0.00	NA
Trans	-0.24	0.15	-1.60	0.11	14.71
HighTrans	0.59	0.15	3.94	0.00	14.15
Checks	0.07	0.05	1.53	0.13	1.47
Years	0.46	0.04	11.58	0.00	1.02

- *Trans* has the highest VIF

[USING *VIF* TO REMOVE COLLINEARITY]

Conclusion:

[Using *VIF* to remove collinearity]

Conclusion:

Omitting one variable (*Trans:*)

[USING *VIF* TO REMOVE COLLINEARITY]

Conclusion:

Omitting one variable (*Trans:*)

	Est.	S.E.	t val.	p	VIF
(Intercept)	5.93	0.04	149.79	0.00	NA
HighTrans	0.36	0.05	7.69	0.00	1.40
Checks	0.06	0.05	1.24	0.22	1.41
Years	0.46	0.04	11.63	0.00	1.02

[USING *VIF* TO REMOVE COLLINEARITY]

Conclusion:

Omitting one variable (*Trans:*)

	Est.	S.E.	t val.	p	VIF
(Intercept)	5.93	0.04	149.79	0.00	NA
HighTrans	0.36	0.05	7.69	0.00	1.40
Checks	0.06	0.05	1.24	0.22	1.41
Years	0.46	0.04	11.63	0.00	1.02

- does not change the fit of the model

[Using *VIF* to remove collinearity]

Conclusion:

Omitting one variable (*Trans:*)

	Est.	S.E.	t val.	p	VIF
(Intercept)	5.93	0.04	149.79	0.00	NA
HighTrans	0.36	0.05	7.69	0.00	1.40
Checks	0.06	0.05	1.24	0.22	1.41
Years	0.46	0.04	11.63	0.00	1.02

- does not change the fit of the model
- does not change the coefficients of the uncorrelated regressors

[USING *VIF* TO REMOVE COLLINEARITY]

Conclusion:

Omitting one variable (*Trans:*)

	Est.	S.E.	t val.	p	VIF
(Intercept)	5.93	0.04	149.79	0.00	NA
HighTrans	0.36	0.05	7.69	0.00	1.40
Checks	0.06	0.05	1.24	0.22	1.41
Years	0.46	0.04	11.63	0.00	1.02

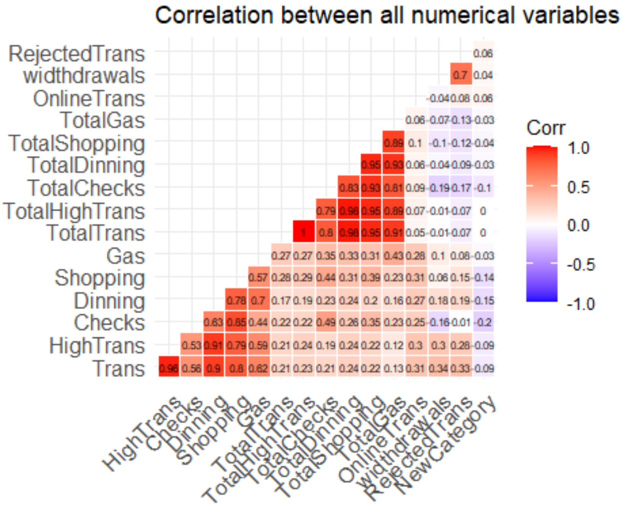
- does not change the fit of the model
- does not change the coefficients of the uncorrelated regressors
- reduces all the *VIFs*

[Real life Example]

In real life, one may have many variables

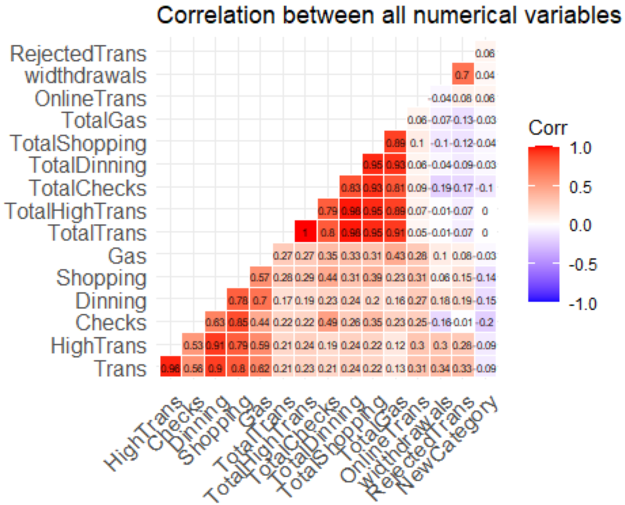
[Real life Example]

In real life, one may have many variables



[Real life Example]

In real life, one may have many variables



[Automatic selection of regressors]

Classic (but still alive) methods based on the variations of RSS

[Automatic selection of regressors]

Classic (but still alive) methods based on the variations of RSS

- Automatic Forward selection

[Automatic selection of regressors]

Classic (but still alive) methods based on the variations of RSS

- Automatic Forward selection
- Automatic Backward selection

[Automatic selection of regressors]

Classic (but still alive) methods based on the variations of RSS

- Automatic Forward selection
- Automatic Backward selection
- Stepwise selection

[Automatic selection of regressors]

Classic (but still alive) methods based on the variations of RSS

- Automatic Forward selection
- Automatic Backward selection
- Stepwise selection

Remark:

[AUTOMATIC SELECTION OF REGRESSORS]

Classic (but still alive) methods based on the variations of RSS

- Automatic Forward selection
- Automatic Backward selection
- Stepwise selection

Remark:

- The optimal number of regressors is unknown!

[AUTOMATIC SELECTION OF REGRESSORS]

Classic (but still alive) methods based on the variations of RSS

- Automatic Forward selection
- Automatic Backward selection
- Stepwise selection

Remark:

- The optimal number of regressors is unknown!

↪ *The number of possible combinations with k regressors is* 2^k

[Automatic selection of regressors]

Classic (but still alive) methods based on the variations of RSS

- Automatic Forward selection
- Automatic Backward selection
- Stepwise selection

Remark:

- The optimal number of regressors is unknown!

↪ *The number of possible combinations with k regressors is* 2^k

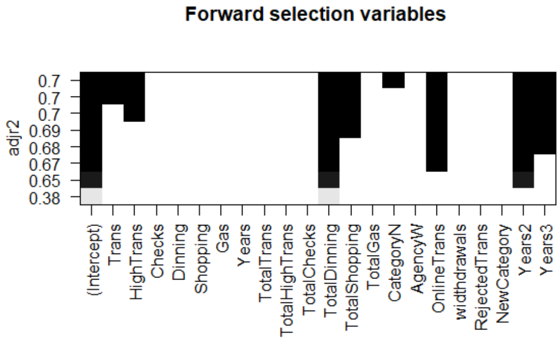
- Compute the optimal nb of regressors before testing which regressors to include with Cross Validation

[Application on an Example]

To reduce the computational burden we restrict our choice to 8 variables in the final regression.

[Application on an Example]

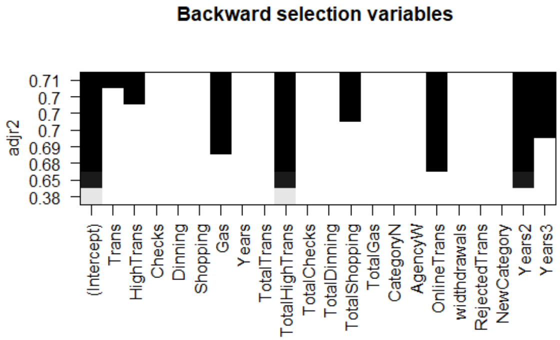
To reduce the computational burden we restrict our choice to 8 variables in the final regression.



Visual representation of variables used (Forward)

[Application on an Example]

To reduce the computational burden we restrict our choice to 8 variables in the final regression.



Visual representation of variables used (Backward)

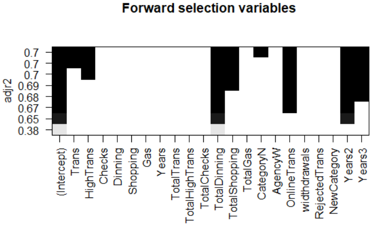
[Criterions for Number of regressors]

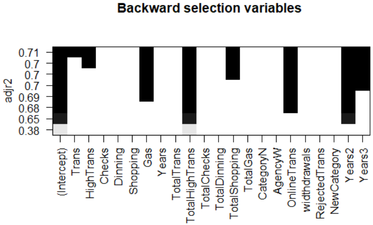
[Criterions for Number of regressors]

- Forward & Backward selection provide different solutions:

[Criterions for Number of regressors]

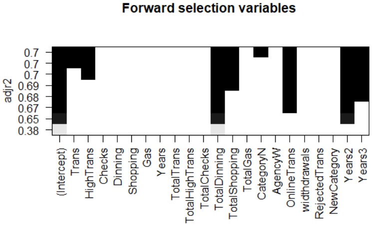
- Forward & Backward selection provide different solutions:

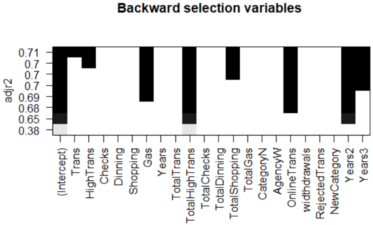




[Criterions for Number of regressors]

- Forward & Backward selection provide different solutions:





- *Great Need for Criteria*

[Automatic selection of regressors]

3 main criteria, based on *RSSs* are found in the literature

[AUTOMATIC SELECTION OF REGRESSORS]

3 main criteria, based on *RSSs* are found in the literature

- Mallow's **Cp** $= \frac{RSS_p}{n} + \frac{2(p+1)}{n} \frac{RSS_k}{n-k-1}$

[AUTOMATIC SELECTION OF REGRESSORS]

3 main criteria, based on $RSSs$ are found in the literature

- Mallow's **Cp** $= \frac{RSS_p}{n} + \frac{2(p+1)}{n}\frac{RSS_k}{n-k-1}$
- Akaike Information Criterion (**AIC**) $\propto C_p$ for linear regression

[AUTOMATIC SELECTION OF REGRESSORS]

3 main criteria, based on $RSSs$ are found in the literature

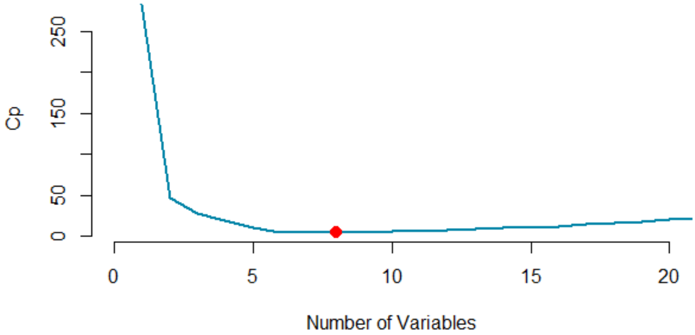
- Mallow's **Cp** $= \frac{RSS_p}{n} + \frac{2(p+1)}{n} \frac{RSS_k}{n-k-1}$
- Akaike Information Criterion (**AIC**) $\propto C_p$ for linear regression
- Bayesian Information Criterion (**BIC**):

$$BIC \propto \frac{RSS_p}{n} + \frac{(p+1)\log n}{n} \frac{RSS_k}{n-k-1}$$

[Application on the Example]

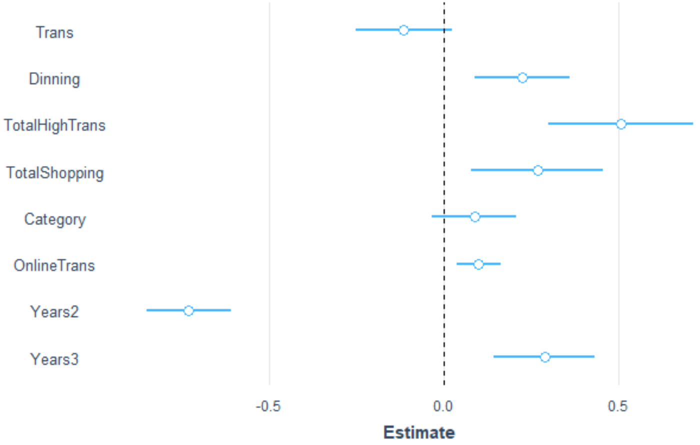
[Application on the Example]

Selection using Mallow's C_p ($\rightarrow$ 8 variables)



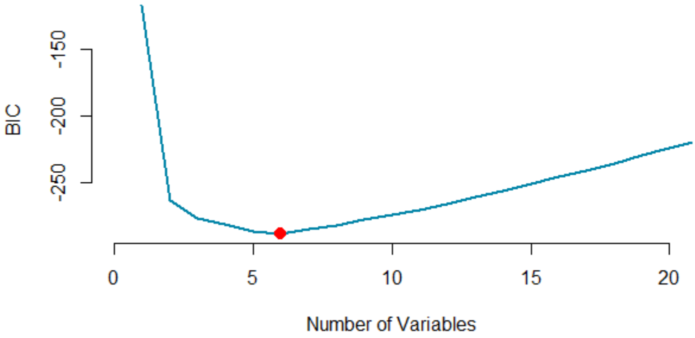
[Application on the Example]

Selection using Mallow's C_p ($\to$ 8 variables)



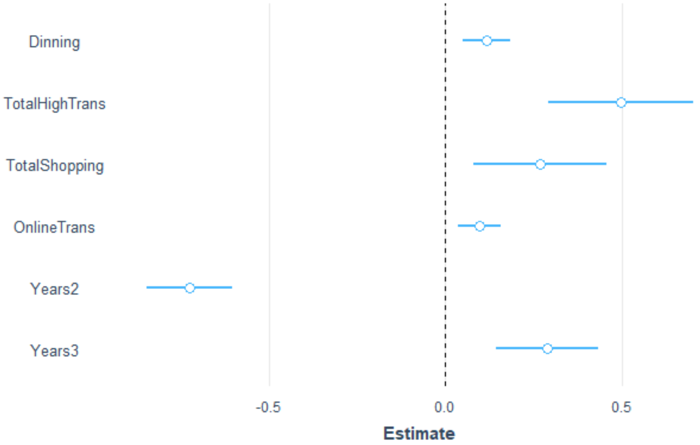
[APPLICATION ON THE EXAMPLE]

Selection using BIC ($\rightarrow$ 6 variables)



[Application on the Example]

Selection using BIC ($\rightarrow$ 6 variables)

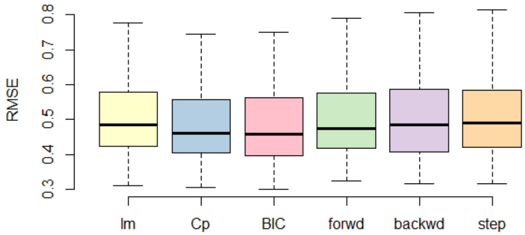


[Cross Validation]

Now we can use cross Validation on all models

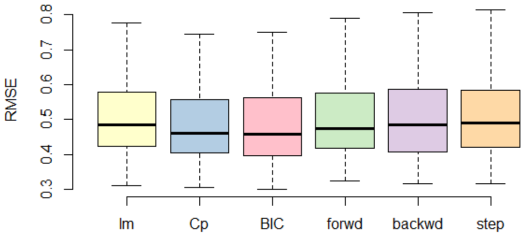
[Cross Validation]

Now we can use cross Validation on all models



[Cross Validation]

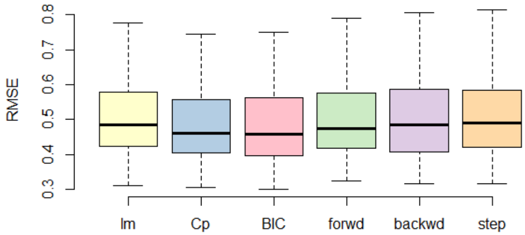
Now we can use cross Validation on all models



- The model selected with C_p is doing the best job

[Cross Validation]

Now we can use cross Validation on all models

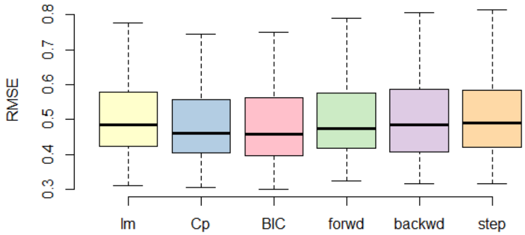


- The model selected with C_p is doing the best job

↪ **Technical issue**: CV with stepwise selects different p-models for each K-fold

[Cross Validation]

Now we can use cross Validation on all models



- The model selected with C_p is doing the best job

↪ **Technical issue**: CV with stepwise selects different p-models for each K-fold

- Modern methods are available...

[Penalization methods]

[Penalization methods]

- *Many* problems when *many* variables are available:

[Penalization methods]

- *Many* problems when *many* variables are available:
  - Multi-collinearity

[Penalization methods]

- *Many* problems when *many* variables are available:
  - Multi-collinearity
  - Model complexity

[Penalization methods]

- *Many* problems when *many* variables are available:
  - Multi-collinearity
  - Model complexity
  - High variance of the estimator

[Penalization methods]

- *Many* problems when *many* variables are available:
  - Multi-collinearity
  - Model complexity
  - High variance of the estimator

$\hookrightarrow$ Methods *"penalizing"* model complexity (over fitting)

[Penalization methods]

- *Many* problems when *many* variables are available:
  - Multi-collinearity
  - Model complexity
  - High variance of the estimator

$\hookrightarrow$ Methods *"penalizing"* model complexity (over fitting)

- Have intensionally a **higher bias** and a **lower variance**

[Penalization methods]

- *Many* problems when *many* variables are available:
  - Multi-collinearity
  - Model complexity
  - High variance of the estimator

↪ Methods *"penalizing"* model complexity (over fitting)

- Have intensionally a **higher bias** and a **lower variance**
- Solutions of a *penalized least-squares problem*

$$\min_{\beta} \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \beta_0 - x_i'\beta\right)^2 + \lambda \cdot J(\beta_1, \cdots, \beta_k)$$

[Penalization methods]

- *Many* problems when *many* variables are available:
  - Multi-collinearity
  - Model complexity
  - High variance of the estimator

↪ Methods *"penalizing"* model complexity (over fitting)

- Have intensionally a **higher bias** and a **lower variance**
- Solutions of a *penalized least-squares problem*

$$\min_{\beta} \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \beta_0 - x_i'\beta\right)^2 + \lambda \cdot J(\beta_1, \cdots, \beta_k)$$

λ is a *hyper parameter*, $J(\cdot)$ is the penalization function

[Penalization methods]

- *Many* problems when *many* variables are available:
  - Multi-collinearity
  - Model complexity
  - High variance of the estimator

$\hookrightarrow$ Methods *"penalizing"* model complexity (over fitting)

- Have intensionally a **higher bias** and a **lower variance**
- Solutions of a *penalized least-squares problem*

$$\min_{\beta} \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \beta_0 - x_i'\beta\right)^2 + \lambda \cdot J(\beta_1, \cdots, \beta_k)$$

λ is a *hyper parameter*, $J(\cdot)$ is the penalization function

NB: *If* $\lambda = 0$ *the regression is just the classic OLS estimator*

[Penalization methods: Ridge regression]

Ridge regression is the solution of:

$$\min_{\beta} \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \beta_0 - x_i'\beta\right)^2 + \lambda \frac{\|\beta\|^2}{2}$$

[Penalization methods: Ridge regression]

Ridge regression is the solution of:

$$\min_{\beta} \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \beta_0 - x_i' \beta\right)^2 + \lambda \frac{\|\beta\|^2}{2}$$

- Ridge regression shrinks parameters towards zero and thus avoids too large parameters

[Penalization methods: Ridge regression]

Ridge regression is the solution of:

$$\min_{\beta} \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \beta_0 - x_i' \beta\right)^2 + \lambda \frac{\|\beta\|^2}{2}$$

- Ridge regression shrinks parameters towards zero and thus avoids too large parameters
- The solution is **biased** intensionally to reduce variance

[Penalization methods: Ridge regression]

Ridge regression is the solution of:

$$\min_{\beta} \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \beta_0 - x_i'\beta\right)^2 + \lambda \frac{\|\beta\|^2}{2}$$

- Ridge regression shrinks parameters towards zero and thus avoids too large parameters
- The solution is **biased** intensionally to reduce variance
- **Penalization** methods prevent these problems by *"penalizing"* model complexity

[Penalization methods: Ridge regression]

Ridge regression is the solution of:

$$\min_{\beta} \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \beta_0 - x_i'\beta\right)^2 + \lambda \frac{\|\beta\|^2}{2}$$

- Ridge regression shrinks parameters towards zero and thus avoids too large parameters
- The solution is **biased** intensionally to reduce variance
- **Penalization** methods prevent these problems by *"penalizing"* model complexity
- It is important to **center and scale** each of the x to ensure the comparability of βs

[Penalization methods: Ridge regression]

Ridge regression is the solution of:

$$\min_{\beta} \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \beta_0 - x_i'\beta\right)^2 + \lambda \frac{\|\beta\|^2}{2}$$

[Penalization methods: Ridge regression]

Ridge regression is the solution of:

$$\min_{\beta} \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \beta_0 - x_i'\beta\right)^2 + \lambda \frac{\|\beta\|^2}{2}$$

- The solution (ridge estimator) is $\widehat{\beta_R} = \left(\frac{X'X}{n} + \lambda I\right)^{-1} \frac{X'y}{n}$

[Penalization methods: Ridge regression]

Ridge regression is the solution of:

$$\min_{\beta} \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \beta_0 - x_i'\beta\right)^2 + \lambda \frac{\|\beta\|^2}{2}$$

- The solution (ridge estimator) is $\widehat{\beta_R} = \left(\frac{X'X}{n} + \lambda I\right)^{-1} \frac{X'y}{n}$
- To compare with OLS estimator $\widehat{\beta_{OLS}} = \left(\frac{X'X}{n}\right)^{-1} \frac{X'y}{n}$

[Penalization methods: Ridge regression]

Ridge regression is the solution of:

$$\min_{\beta} \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \beta_0 - x_i'\beta\right)^2 + \lambda \frac{\|\beta\|^2}{2}$$

- The solution (ridge estimator) is $\widehat{\beta_R} = \left(\frac{X'X}{n} + \lambda I\right)^{-1} \frac{X'y}{n}$
- To compare with OLS estimator $\widehat{\beta_{OLS}} = \left(\frac{X'X}{n}\right)^{-1} \frac{X'y}{n}$

↪ When there is collinearity, $X'X$ cannot be inverted, while if $\lambda > 0$ the matrix $(X'X + \lambda I)$ is invertible

[Ridge regression in Practice]

How to choose λ?

[Ridge regression in Practice]

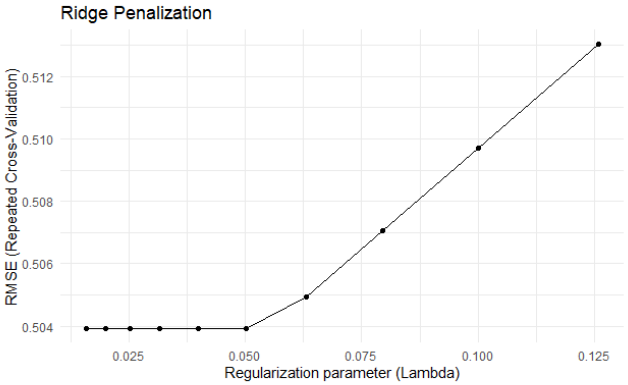
How to choose λ?

- We compute the (CV-averaged) *RMSE* for many values of λ

[Ridge regression in Practice]

How to choose λ?

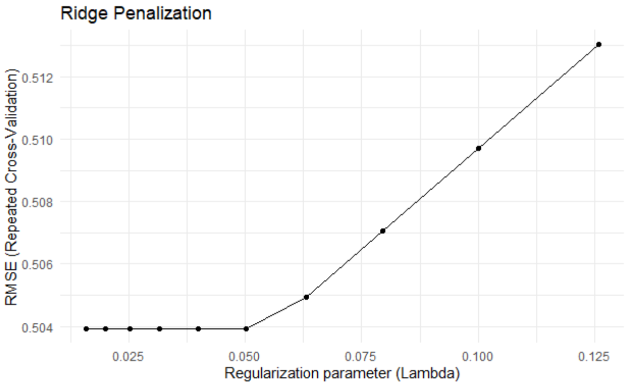
- We compute the (CV-averaged) *RMSE* for many values of λ



[Ridge regression in Practice]

How to choose λ?

- We compute the (CV-averaged) *RMSE* for many values of λ



Remark: λ is typically small, and we usually select it on the log scale.

[Penalization methods: LASSO]

LASSO or *Least Absolute Shrinkage and Selection Operator*, is another common penalization method, is the solution of:

$$\min_{\beta} \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \beta_0 - x_i'\beta\right)^2 + \lambda|\beta| \qquad |\beta| = \sum_{j=1}^{k} |\beta_j|$$

[Penalization methods: LASSO]

LASSO or *Least Absolute Shrinkage and Selection Operator*, is another common penalization method, is the solution of:

$$\min_{\beta} \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \beta_0 - x_i'\beta\right)^2 + \lambda|\beta| \qquad |\beta| = \sum_{j=1}^{k} |\beta_j|$$

$\hookrightarrow$ If $\lambda = 0$, we obtain OLS. If $\lambda = \infty$, all parameters are zero.

[Penalization methods: LASSO]

LASSO or *Least Absolute Shrinkage and Selection Operator*, is another common penalization method, is the solution of:

$$\min_{\beta} \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \beta_0 - x_i'\beta\right)^2 + \lambda|\beta| \qquad |\beta| = \sum_{j=1}^{k} |\beta_j|$$

↪ If $\lambda = 0$, we obtain OLS. If $\lambda = \infty$, all parameters are zero.

- Lasso does automatic variable selection: if λ is large enough, the solution put some parameters to zero

[Penalization methods: LASSO]

LASSO or *Least Absolute Shrinkage and Selection Operator*, is another common penalization method, is the solution of:

$$\min_{\beta} \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \beta_0 - x_i'\beta\right)^2 + \lambda|\beta| \qquad |\beta| = \sum_{j=1}^{k} |\beta_j|$$

$\hookrightarrow$ If $\lambda = 0$, we obtain OLS. If $\lambda = \infty$, all parameters are zero.

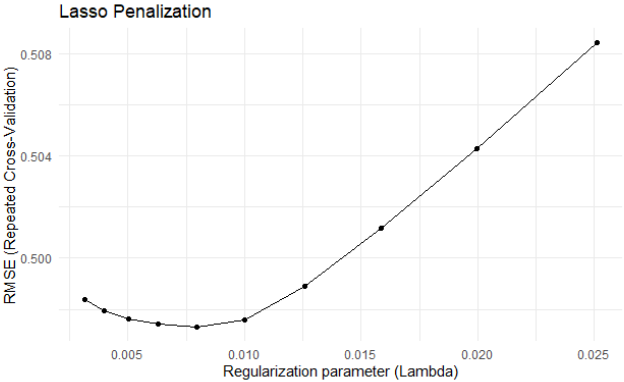
- Lasso does automatic variable selection: if λ is large enough, the solution put some parameters to zero
- It is important to **center and scale** each of the x to ensure the comparability of βs

[LASSO in Practice]

We compute the (CV-averaged) *RMSE* for many values of λ

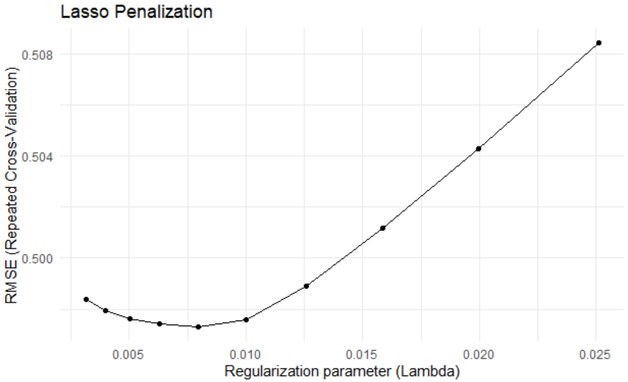
[LASSO IN PRACTICE]

We compute the (CV-averaged) *RMSE* for many values of λ



[LASSO in Practice]

We compute the (CV-averaged) *RMSE* for many values of λ



Remark: λ is typically small, and we usually select it on the log scale.

[Penalization methods: Elastic Net]

Elastic Net combines both Lasso and Ridge regression:

$$\min_{\beta} \frac{1}{2n} \sum_{i=1}^{n} (y_i - \beta_0 - x_i'\beta)^2 + \lambda \left((1-\alpha)\frac{\|\beta\|^2}{2} + \alpha|\beta| \right)$$

[Penalization methods: Elastic Net]

Elastic Net combines both Lasso and Ridge regression:

$$\min_{\beta} \frac{1}{2n} \sum_{i=1}^{n} (y_i - \beta_0 - x_i'\beta)^2 + \lambda \left((1-\alpha)\frac{||\beta||^2}{2} + \alpha|\beta| \right)$$

- If $\alpha = 1$ we have the Lasso estimator, if $\alpha = 0$, the Ridge regression.

[Penalization methods: Elastic Net]

Elastic Net combines both Lasso and Ridge regression:

$$\min_{\beta} \frac{1}{2n} \sum_{i=1}^{n} (y_i - \beta_0 - x_i'\beta)^2 + \lambda \left((1-\alpha) \frac{\|\beta\|^2}{2} + \alpha|\beta| \right)$$

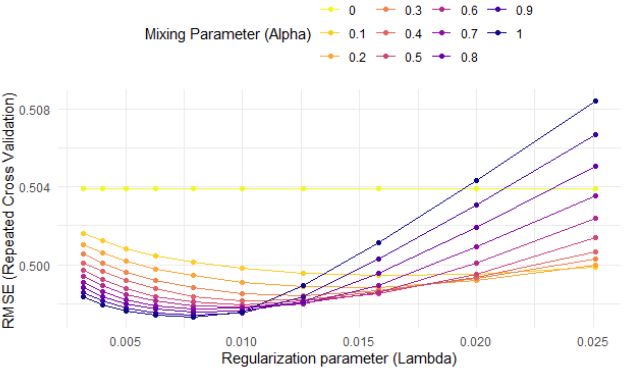
- If $\alpha = 1$ we have the Lasso estimator, if $\alpha = 0$, the Ridge regression.
- Through α we balance variable selection (Lasso) and coefficient reduction (Ridge)

[Elastic Net in Practice]

Compute the (CV-averaged) *RMSE* on a grid of $(\lambda\ ,\ \alpha)$

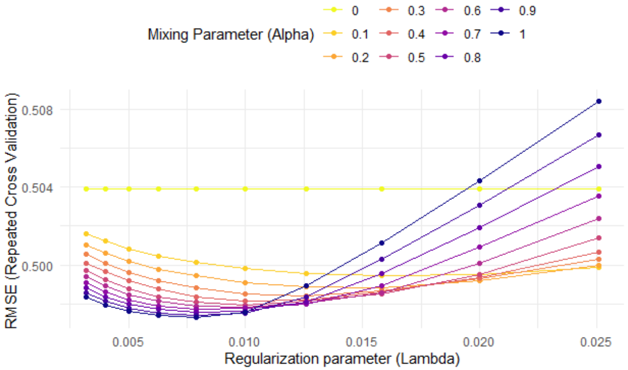
[Elastic Net in Practice]

Compute the (CV-averaged) *RMSE* on a grid of $(\lambda\ ,\ \alpha)$



[Elastic Net in Practice]

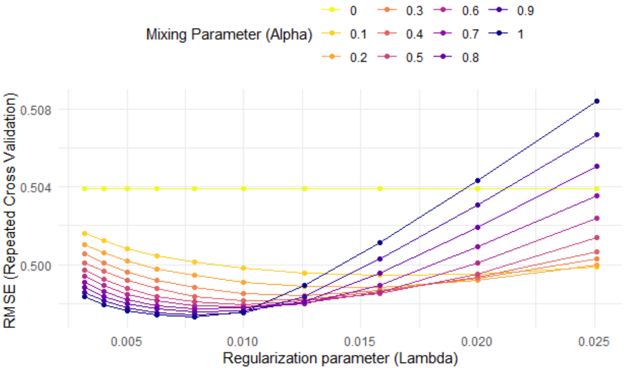
Compute the (CV-averaged) *RMSE* on a grid of (λ , α)



$\hookrightarrow$ Optimal value: $\alpha^* = 1$ & $\lambda = 0.008 \hookrightarrow$ Lasso estimator.

[Elastic Net in Practice]

Compute the (CV-averaged) *RMSE* on a grid of $(\lambda\ ,\ \alpha)$



$\hookrightarrow$ Optimal value: $\alpha^* = 1$ & $\lambda = 0.008 \hookrightarrow$ Lasso estimator.

- Elastic net *encompass* both Ridge and Lasso estimators.

[Best Model?]

In Machine Learning, focus on prediction (RMSE)

[Best Model?]

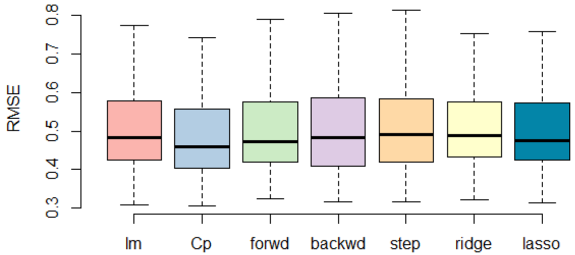
In Machine Learning, focus on prediction (RMSE)

- Cross-Validation performance (RMSE):

[Best Model?]

In Machine Learning, focus on prediction (RMSE)

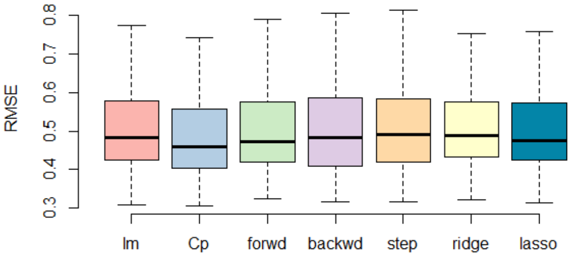
- Cross-Validation performance (RMSE):



[Best Model?]

In Machine Learning, focus on prediction (RMSE)

- Cross-Validation performance (RMSE):

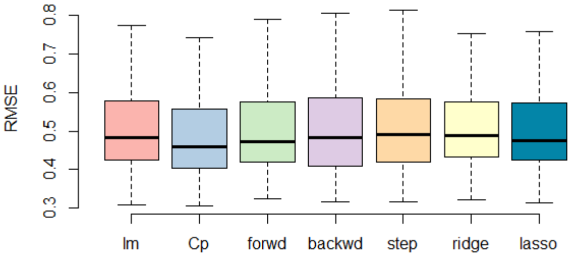


- LASSO is probably the best (lower RMSE)

[Best Model?]

In Machine Learning, focus on prediction (RMSE)

- Cross-Validation performance (RMSE):



- LASSO is probably the best (lower RMSE)
- Model selected by Mallow's C_p (8 regressors) is almost as good!

[TAKEAWAYS]

- In linear regression, scaling allows to compare coefficients and to measure variable importance.

[Takeaways]

- In linear regression, scaling allows to compare coefficients and to measure variable importance.
- Multi-collinearity should be investigated beforehand.

[TAKEAWAYS]

- In linear regression, scaling allows to compare coefficients and to measure variable importance.
- Multi-collinearity should be investigated beforehand.
- Mallows C_p is a simple method to select regressors

[Takeaways]

- In linear regression, scaling allows to compare coefficients and to measure variable importance.
- Multi-collinearity should be investigated beforehand.
- Mallows C_p is a simple method to select regressors
- Stepwise methods with CV have three drawbacks:

[TAKEAWAYS]

- In linear regression, scaling allows to compare coefficients and to measure variable importance.
- Multi-collinearity should be investigated beforehand.
- Mallows C_p is a simple method to select regressors
- Stepwise methods with CV have three drawbacks:
  - they do not necessarily select the "best" model

[TAKEAWAYS]

- In linear regression, scaling allows to compare coefficients and to measure variable importance.
- Multi-collinearity should be investigated beforehand.
- Mallows C_p is a simple method to select regressors
- Stepwise methods with CV have three drawbacks:
  - they do not necessarily select the "best" model
  - the choice of variables can be sensitive to the number of repetitions

[Takeaways]

- In linear regression, scaling allows to compare coefficients and to measure variable importance.
- Multi-collinearity should be investigated beforehand.
- Mallows C_p is a simple method to select regressors
- Stepwise methods with CV have three drawbacks:
  - they do not necessarily select the "best" model
  - the choice of variables can be sensitive to the number of repetitions
  - Variable selection/elimination is done variable by variable (no interactions)

[Takeaways]

- Penalized least-squares methods are useful for dealing with multicollinearity or with a large number of regressors.

[TAKEAWAYS]

- Penalized least-squares methods are useful for dealing with multicollinearity or with a large number of regressors.
- All solutions of a *penalized least-squares problem*

$$\min_{\beta} \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \beta_0 - x_i'\beta\right)^2 + \lambda \cdot J_\alpha(\beta_1, \cdots, \beta_k)$$

[Takeaways]

- Penalized least-squares methods are useful for dealing with multicollinearity or with a large number of regressors.
- All solutions of a *penalized least-squares problem*

$$\min_{\beta} \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \beta_0 - x_i'\beta\right)^2 + \lambda \cdot J_\alpha(\beta_1, \cdots, \beta_k)$$

λ is a *hyperparameter*, $J_\alpha(\cdot)$ is the penalization function

[TAKEAWAYS]

- Penalized least-squares methods are useful for dealing with multicollinearity or with a large number of regressors.
- All solutions of a *penalized least-squares problem*

$$\min_{\beta} \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \beta_0 - x_i'\beta\right)^2 + \lambda \cdot J_\alpha(\beta_1, \cdots, \beta_k)$$

  λ is a *hyperparameter*, $J_\alpha(\cdot)$ is the penalization function
- It is important to **center and scale** each of the x to ensure the comparability of βs

[Takeaways]

- Penalized least-squares methods are useful for dealing with multicollinearity or with a large number of regressors.
- All solutions of a *penalized least-squares problem*

$$\min_{\beta} \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \beta_0 - x_i'\beta\right)^2 + \lambda \cdot J_\alpha(\beta_1, \cdots, \beta_k)$$

  λ is a *hyperparameter*, $J_\alpha(\cdot)$ is the penalization function
- It is important to **center and scale** each of the x to ensure the comparability of βs
- Selection of *hyper parameters* is based on CV and RMSE

[Takeaways]

- Penalized least-squares methods are useful for dealing with multicollinearity or with a large number of regressors.
- All solutions of a *penalized least-squares problem*

$$\min_{\beta} \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \beta_0 - x_i'\beta\right)^2 + \lambda \cdot J_\alpha(\beta_1, \cdots, \beta_k)$$

  λ is a *hyperparameter*, $J_\alpha(\cdot)$ is the penalization function
- It is important to **center and scale** each of the x to ensure the comparability of βs
- Selection of *hyper parameters* is based on CV and RMSE
- The elastic net "encompass" both Ridge and Lasso estimators.

[Takeaways]

- Penalized least-squares methods are useful for dealing with multicollinearity or with a large number of regressors.
- All solutions of a *penalized least-squares problem*

$$\min_{\beta} \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \beta_0 - x_i'\beta\right)^2 + \lambda \cdot J_\alpha(\beta_1, \cdots, \beta_k)$$

  λ is a *hyperparameter*, $J_\alpha(\cdot)$ is the penalization function
- It is important to **center and scale** each of the x to ensure the comparability of βs
- Selection of *hyper parameters* is based on CV and RMSE
- The elastic net "encompass" both Ridge and Lasso estimators.
- Penalization methods are very popular in practice

[Q&A]

Write your questions in the chat

[Next week]

[Next week]

- **NO Webinar next Thursday**

[Next week]

- **NO Webinar next Thursday**
- Enjoy the activities proposed - Module 4 opens soon

[Next week]

- **NO Webinar next Thursday**
- Enjoy the activities proposed - Module 4 opens soon

$\hookrightarrow$ A personal ML project will be proposed soon!

[Next week]

- **NO Webinar next Thursday**
- Enjoy the activities proposed - Module 4 opens soon

$\hookrightarrow$ A personal ML project will be proposed soon!

- **Continue to post your thoughts on the forums**

[Next week]

- **NO Webinar next Thursday**
- Enjoy the activities proposed - Module 4 opens soon

↪ A personal ML project will be proposed soon!

- **Continue to post your thoughts on the forums**

Have a nice week!