Report 1

LASSO, Ridge and ElasticNet Regression

Klaudia Balcer

12/17/2021

Contents

Task 1 **2**

Theoretical calculations for Ridge Regression 2

Simulations 3

Results 4

Task 2 **7**

Theoretical calculations 7

Simulations 7

Results 8

Task 3 **10**

Simulations 10

Results 11

Task 4 **13**

Simulations a 13

Simulations b 13

Results b 14

Results b 16

Task 5 **18**

Theoretical descritpion 18

Simulation 18

Task 6 **19**

Implementation 20

Results 20

Task 1

In the first task, we will compare the properties of Ordinary Least Squares (OLS) and Ridge regression. We will consider the variance, the bias, and the mean squared error (MSE) of $\hat{\beta}$ in the orthogonal design.

Theoretical calculations for Ridge Regression

1. **Coefficients**

$$\hat{\beta}^{RIDGE} = \frac{1}{1+\gamma}\hat{\beta}^{LS}$$

2. **Optimal γ (in terms of MSE)**

$$\gamma_{opt} = \frac{p\sigma^2}{\|\beta\|^2}$$

3. **MSE**

- general:

$$MSE(\gamma) = \frac{\gamma^2\|\beta\|^2 + p\sigma^2}{(1+\gamma)^2}$$

- for optimal γ:

$$MSE_{opt} = \frac{\|\beta\|^2 p\sigma^2}{\|\beta\|^2 + p\sigma^2}$$

4. **Bias**

- general:

$$\mathbb{E}[\hat{\beta}^{RIDGE} - \beta] = \mathbb{E}\hat{\beta}^{RIDGE} - \beta = \mathbb{E}[\frac{1}{1+\gamma}\hat{\beta}^{LS}] - \beta = \frac{1}{1+\gamma}\mathbb{E}\hat{\beta}^{LS} - \beta = \frac{1}{1+\gamma}\beta - \beta = -\frac{\gamma}{1+\gamma}\beta$$

- for optimal γ:

$$Bias_{opt} = -\frac{p\sigma^2}{p\sigma^2 + \|\beta\|^2}\beta$$

5. **Variance**

- general:

$$Var(\hat{\beta}_i^{RIDGE}) = Var\Big(\frac{1}{1+\gamma}(\beta_i + Z_i)\Big) = \Big(\frac{1}{1+\gamma}\Big)^2 Var(\beta_i + Z_i) = \Big(\frac{1}{1+\gamma}\Big)^2 Var(Z_i) = \Big(\frac{1}{1+\gamma}\Big)^2\sigma^2 = \frac{\sigma^2}{(1+\gamma)^2}$$

- for optimal γ:

$$\frac{\sigma^2\|\beta\|^4}{(p\sigma^2 + \|\beta\|^2)^2}$$

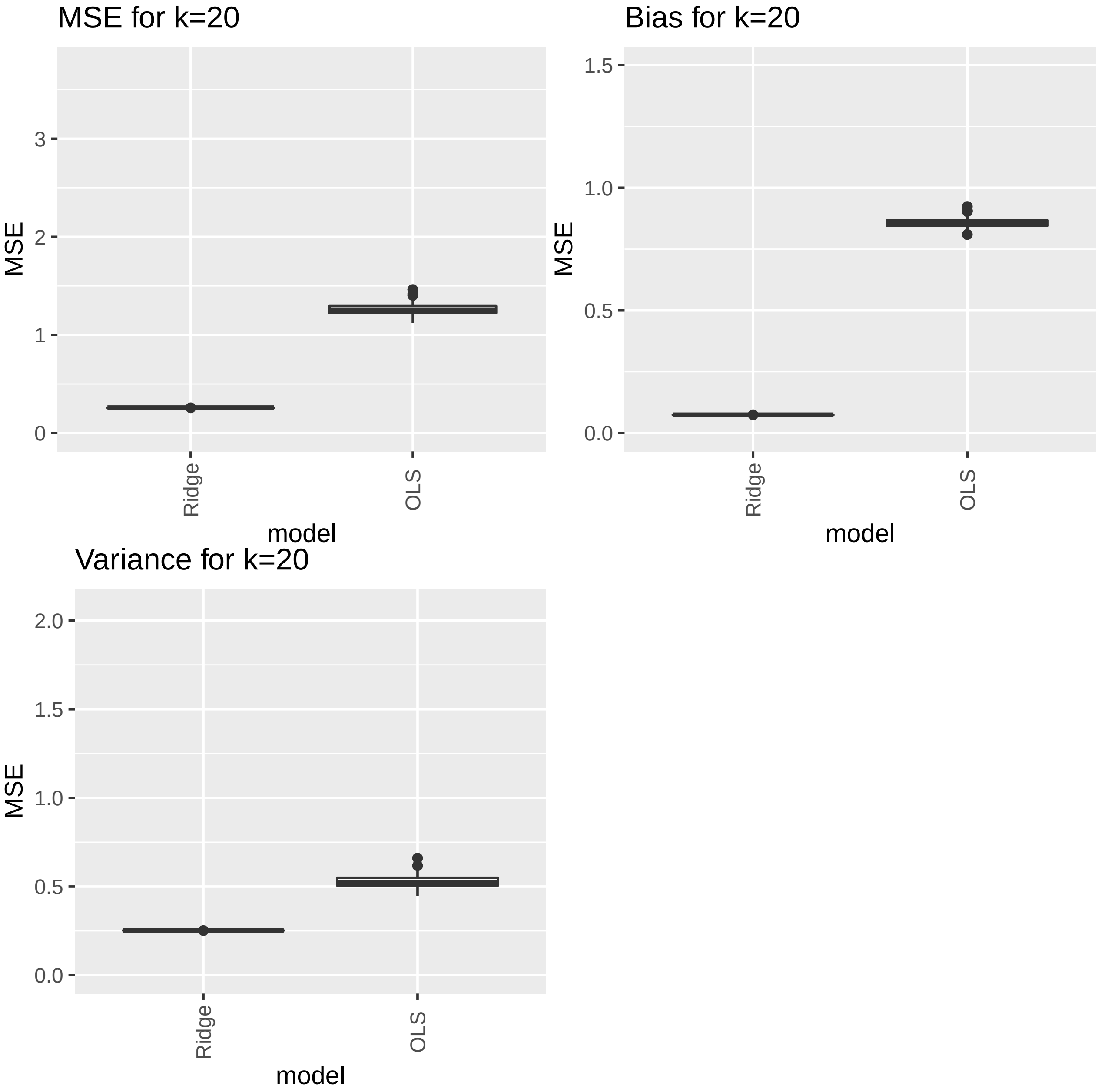
Simulations

We will repeat the following experiment two hundred times:

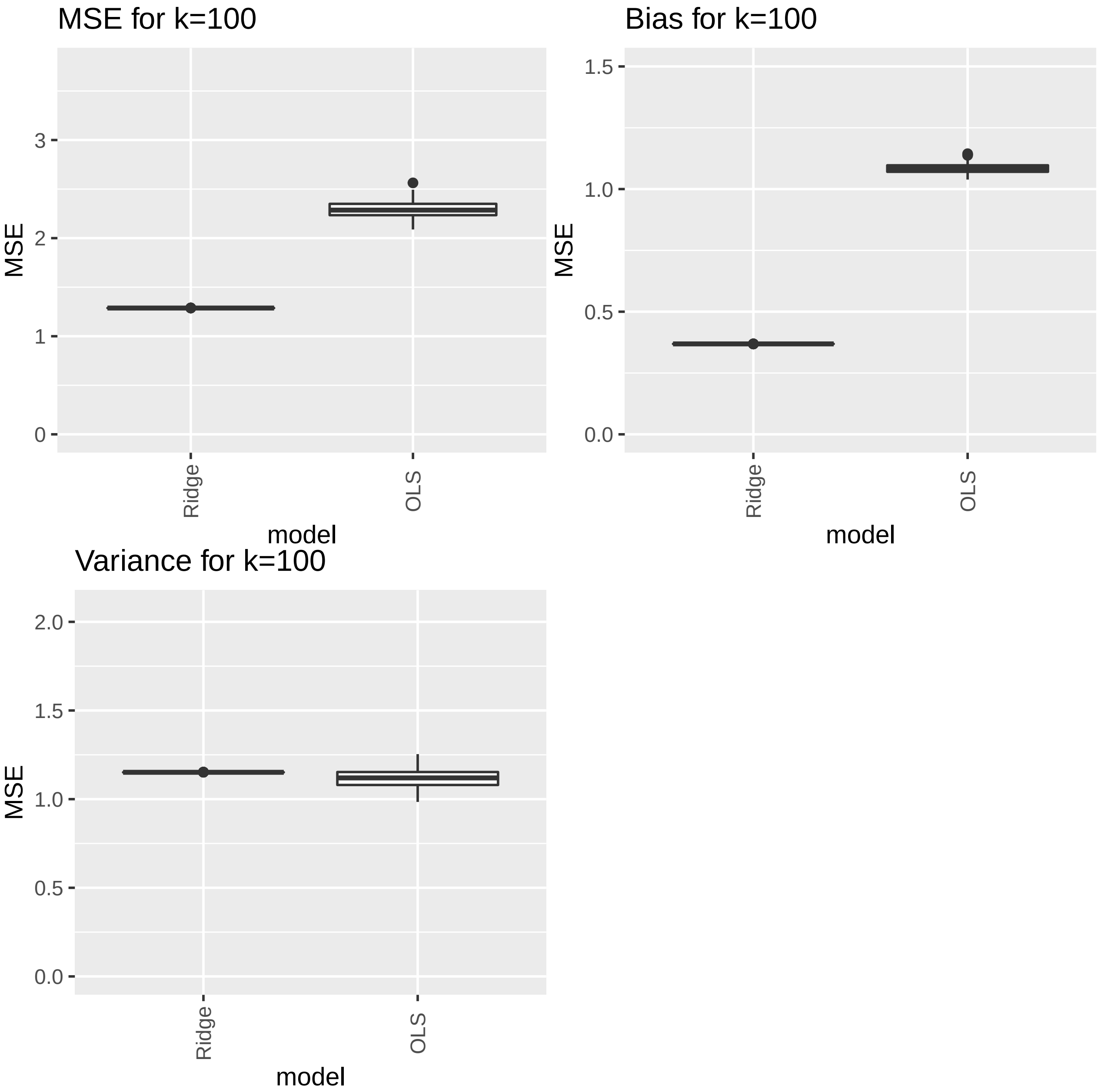
- generate an orthogonal matrix X of size 1000×950, generate an error term vector from standard normal distribution;
- the real coefficients are: $\beta_1, \ldots, \beta_k = 3.5$, $\beta_{k+1}, \ldots, \beta_{950} = 0$ for k equal 20, 100, 200;
- select tuning parameter λ for Ridge Regression to minimize the MSE of $\hat{\beta}$;
- build an OLS and Ridge Regression model;
- evaluate the models by calculating MSE, variance and bias.

Results

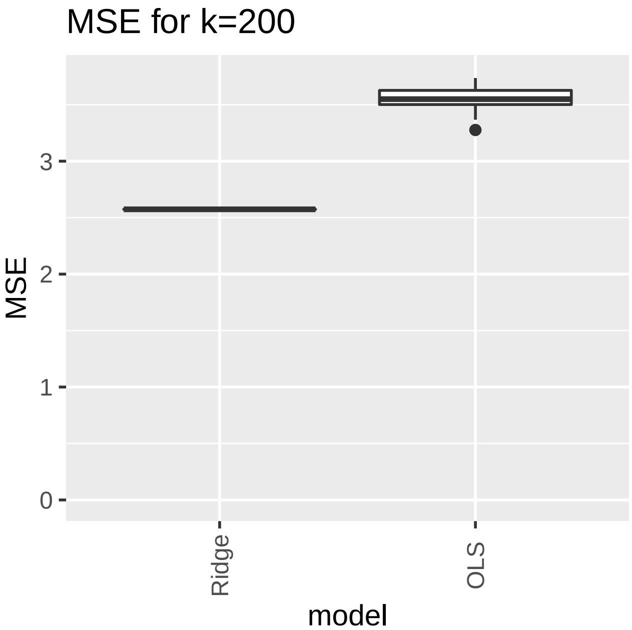
Results for k=20

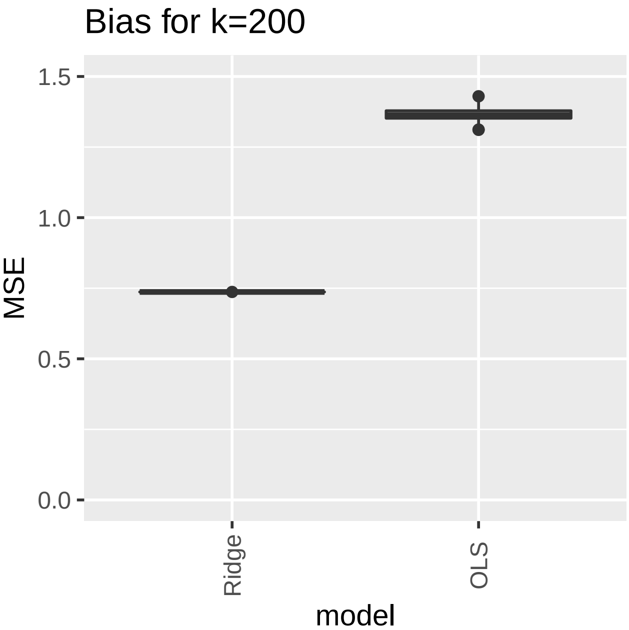


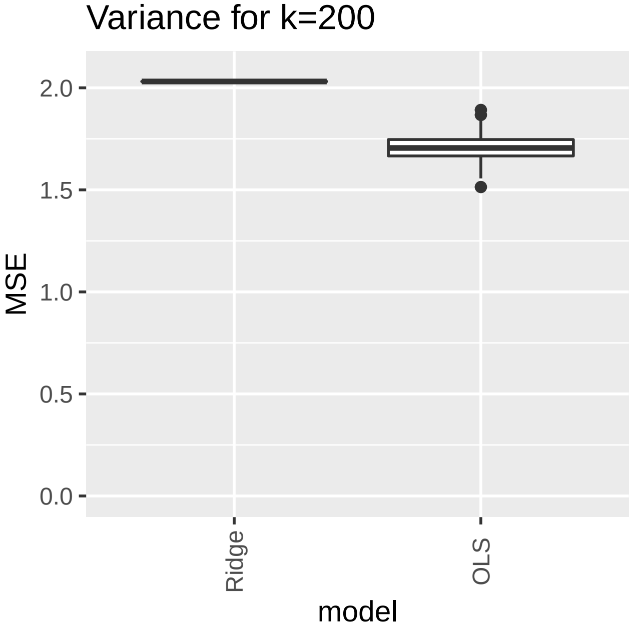
Results for k=100



Results for k=200







Results summary

Table 1: Comparison of OLS and Ridge Regression

	20	100	200
MSE_rr	0.26	1.29	2.57
bias_rr	0.07	0.37	0.74
var_rr	0.25	1.15	2.03
MSE_ols	1.26	2.29	3.58
bias_ols	0.86	1.08	1.37
var_ols	0.53	1.12	1.71

Both methods work better for small number of effects. When the signal is sparse, Ridge outperforms OLS. When the number of significant variables increases, the variance of Ridge also increases. For k=200, Ridge has a greater variance than OLS. However, it has still much smaller bias (and smaller MSE in consequence).

Task 2

In this and the upcoming tasks, we will compare several regression approaches. We will compare the MSE of predictions and coefficients for all those approaches.

Theoretical calculations

Prediction error for ridge regression (using SURE):

$$\widehat{PE} = RSS + 2\sigma^2 \sum_{i=1}^{n} \frac{\lambda_i(X^T X)}{\lambda_i(X^T X) + \gamma}$$

where λ_i are the eigen values of $H = X(X^T X + \gamma I)^{-1} X^T$.

Prediction error for LASSO (using SURE):

$$\widehat{PE} = RSS + 2\sigma^2 \# \mathcal{A}$$

where $\#\mathcal{A}$ is the number of selected variables.

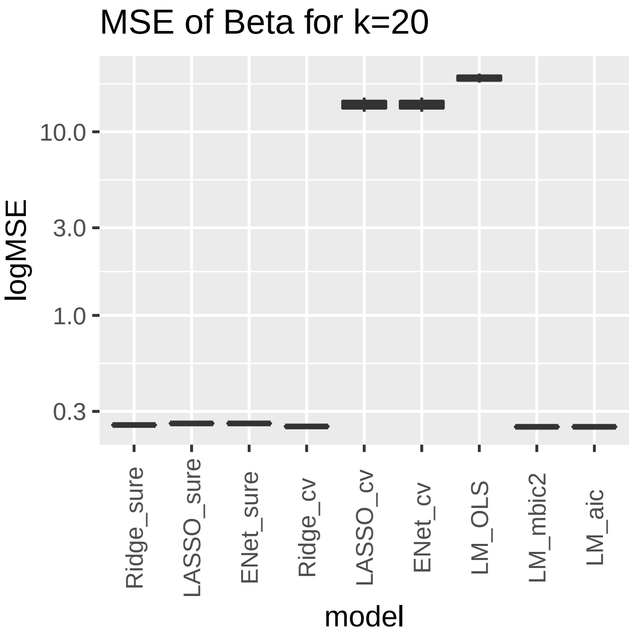
Simulations

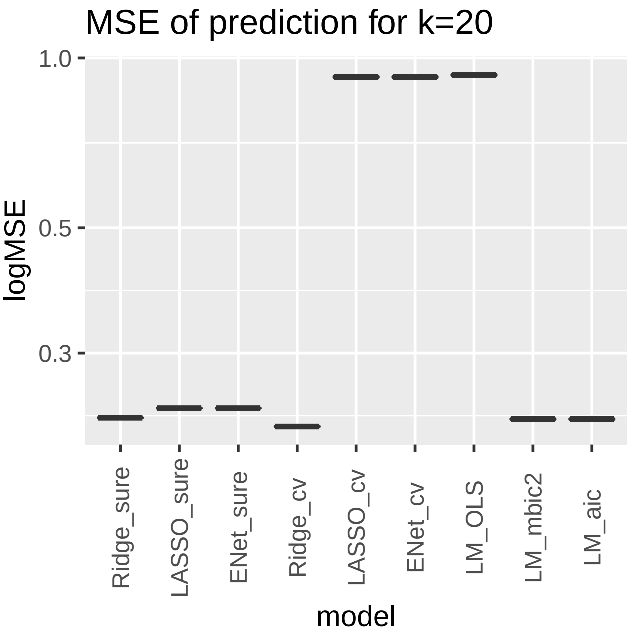
We will repeat the following experiment a hundred times:

- sample matrix X of size 1000×950 from normal distribution $\mathcal{N}(0, \sigma = frac1\sqrt{n})$, generate an error term vector from standard normal distribution;
- the real coefficients are: $\beta_1, \ldots, \beta_k = 3.5$, $\beta_{k+1}, \ldots, \beta_{950} = 0$ for k equal 20, 100, 200;
- build models:
  - LASSO with
    - λ from CV,
    - λ from SURE;
  - Ridge Regression
    - λ from CV,
    - λ from SURE;
  - ElasticNet with $\alpha = 0.5$ and
    - λ from CV,
    - λ from SURE;
  - OLS
    - with all variables,
    - with variables selected by AIC,
    - with variables selected by mBIC2.
- evaluate the models by calculating MSE for $\hat{\beta}$ and $\hat{Y}$.

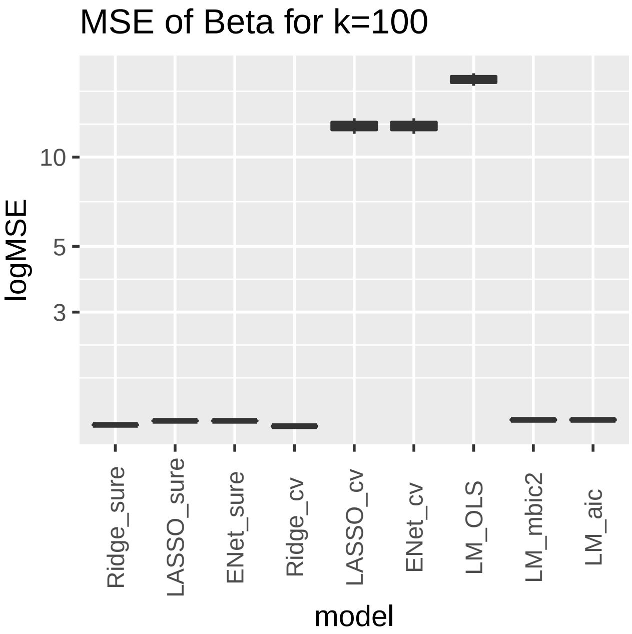
Results

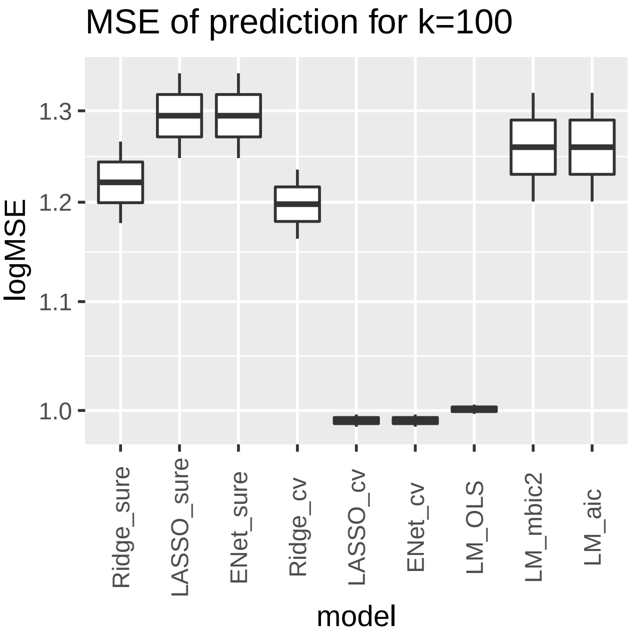
Results for k=20



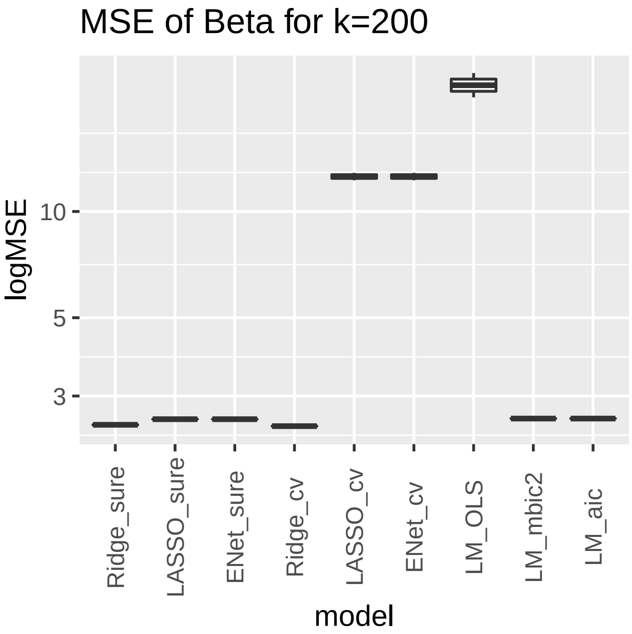


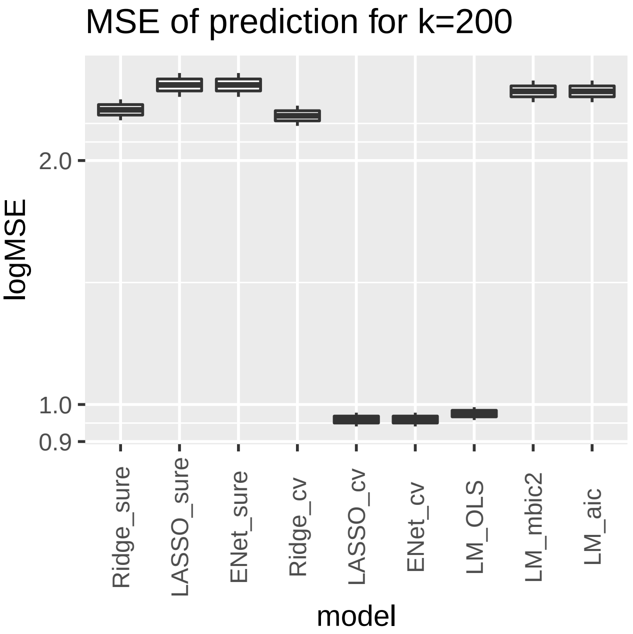
Results for k=100





Results for k=200





Results summary

Table 2: MSE of Beta

	20	100	200
Ridge_sure	0.25	1.25	2.49
LASSO_sure	0.26	1.29	2.58
ENet_sure	0.26	1.29	2.58
Ridge_cv	0.25	1.24	2.46
LASSO_cv	14.11	12.75	12.56
ENet_cv	14.11	12.75	12.56
LM_OLS	19.64	18.28	22.86
LM_mbic2	0.25	1.30	2.59
LM_aic	0.25	1.30	2.59

Table 3: MSE of prediction

	20	100	200
Ridge_sure	0.23	1.22	2.31
LASSO_sure	0.24	1.30	2.48
ENet_sure	0.24	1.30	2.48
Ridge_cv	0.22	1.20	2.27
LASSO_cv	0.93	0.99	0.96
ENet_cv	0.93	0.99	0.96
LM_OLS	0.93	1.00	0.97
LM_mbic2	0.23	1.26	2.44
LM_aic	0.23	1.26	2.44

Although the 10-fold cross-validation does not estimate coefficients precisely, its prediction properties are well and do not depend on k. In most cases, it works at least as well as OLS. When minimizing the Prediction Error Estimator obtained from SURE, the results are similar to those for non-regularised linear models with preselected variables. Results for both of these groups get worse when k increases.

Task 3

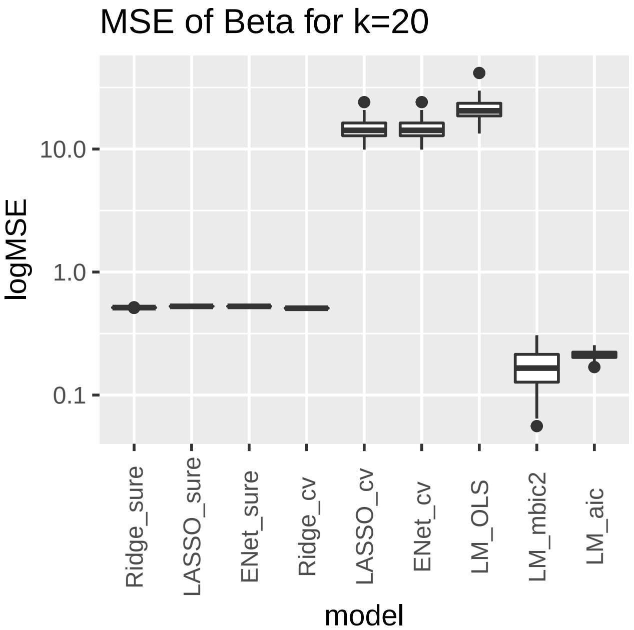
In this task we will repeat the previous experiment for stronger signal.

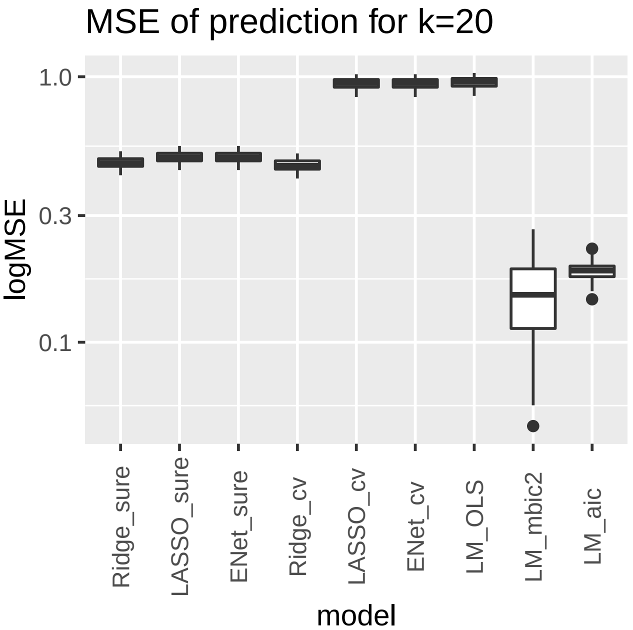
Simulations

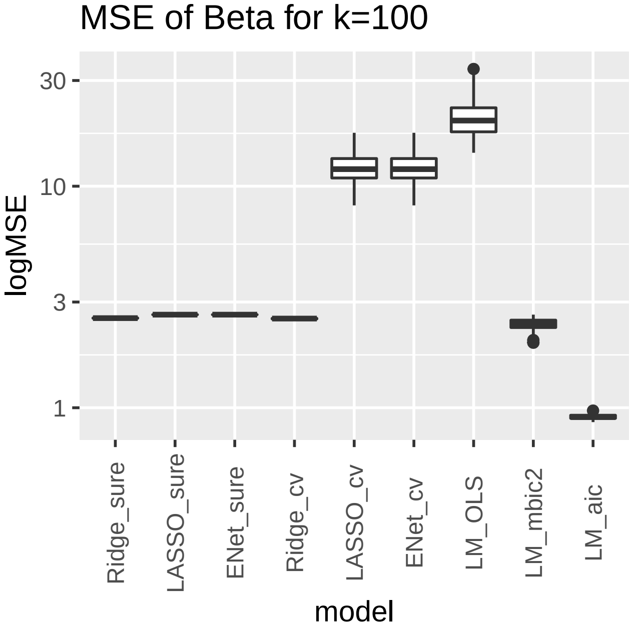
We will repeat the following experiment a hundred times:

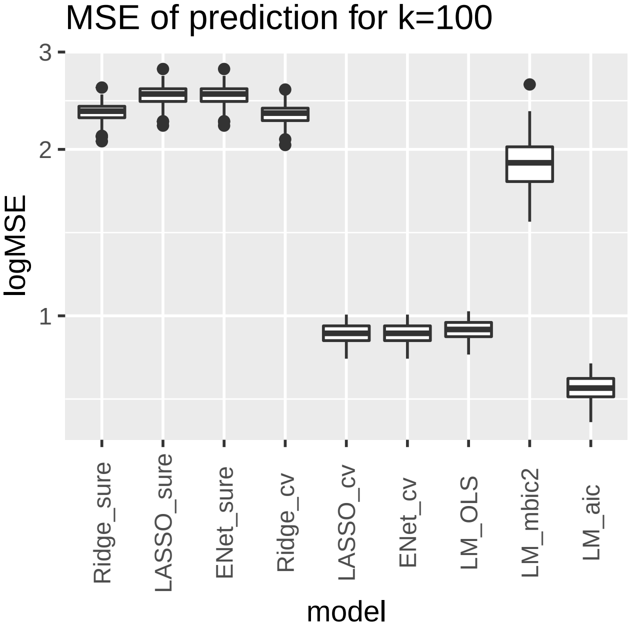
- sample matrix X of size 1000×950 from normal distribution $\mathcal{N}(0, \sigma = frac1\sqrt{n})$, generate an error term vector from standard normal distribution;
- **the real coefficients are:** $\beta_1, \ldots, \beta_k = 5$**,** $\beta_{k+1}, \ldots, \beta_{950} = 0$ **for** k **equal 20, 100, 200;**
- build models:
  - LASSO with
    - λ from CV,
    - λ from SURE;
  - Ridge Regression
    - λ from CV,
    - λ from SURE;
  - ElasticNet with $\alpha = 0.5$ and
    - λ from CV,
    - λ from SURE;
  - OLS
    - with all variables,
    - with variables selected by AIC,
    - with variables selected by mBIC2.
- evaluate the models by calculating MSE for $\hat{\beta}$ and $\hat{Y}$.

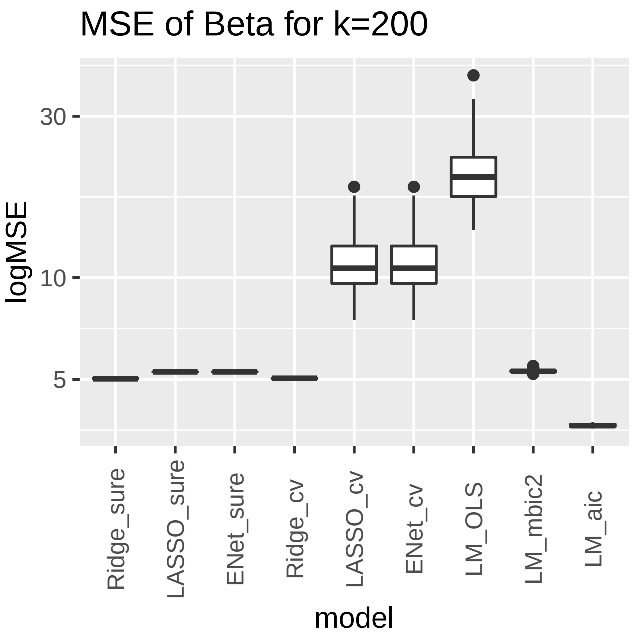
Results


MSE of Beta for k=20
logMSE
10.0
1.0
0.1
Ridge_sure
LASSO_sure
ENet_sure
Ridge_cv
LASSO_cv
ENet_cv
LM_OLS
LM_mbic2
LM_aic
model

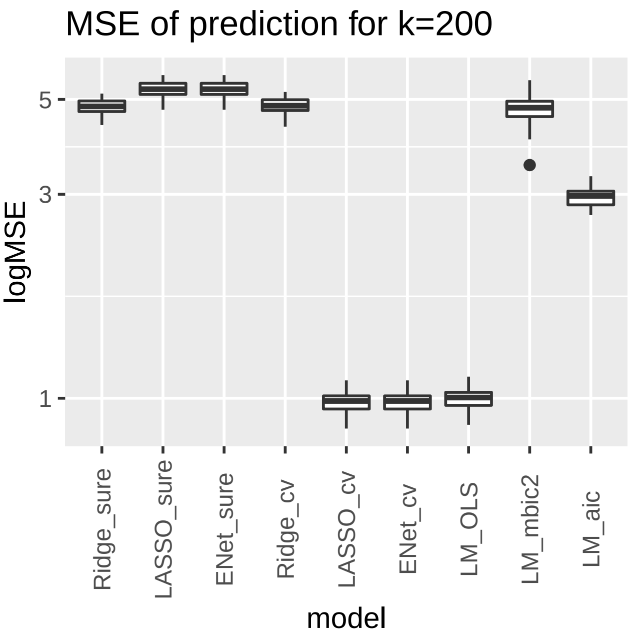

MSE of prediction for k=20
logMSE
1.0
0.3
0.1
Ridge_sure
LASSO_sure
ENet_sure
Ridge_cv
LASSO_cv
ENet_cv
LM_OLS
LM_mbic2
LM_aic
model


MSE of Beta for k=100
logMSE
30
10
3
1
Ridge_sure
LASSO_sure
ENet_sure
Ridge_cv
LASSO_cv
ENet_cv
LM_OLS
LM_mbic2
LM_aic
model


MSE of prediction for k=100
logMSE
3
2
1
Ridge_sure
LASSO_sure
ENet_sure
Ridge_cv
LASSO_cv
ENet_cv
LM_OLS
LM_mbic2
LM_aic
model



MSE of Beta for k=200



MSE of prediction for k=200

Table 4: MSE of Beta

	20	100	200
Ridge_sure	0.51	2.54	5.02
LASSO_sure	0.53	2.63	5.26
ENet_sure	0.53	2.63	5.26
Ridge_cv	0.51	2.53	5.03
LASSO_cv	14.62	12.16	11.11
ENet_cv	14.62	12.16	11.11
LM_OLS	21.20	20.47	20.47
LM_mbic2	0.17	2.38	5.29
LM_aic	0.21	0.91	3.66

Table 5: MSE of prediction

	20	100	200
Ridge_sure	0.48	2.34	4.82
LASSO_sure	0.50	2.51	5.28
ENet_sure	0.50	2.51	5.28
Ridge_cv	0.47	2.32	4.84
LASSO_cv	0.94	0.93	0.98
ENet_cv	0.94	0.93	0.98
LM_OLS	0.95	0.94	1.00
LM_mbic2	0.15	1.90	4.75
LM_aic	0.19	0.74	2.96

OLS with preselected variables got better in sparse cases. All other results became worse.

Task 4

Simulations a

We will repeat the following experiment a hundred times:

- sample matrix X of size 1000×950 from multivariate normal distribution $\mathcal{N}_n(0, \Sigma = 0.5 + 0.5 * \mathbb{1}i = 1)$, generate an error term vector from standard normal distribution;
- the real coefficients are: $\beta_1, \ldots, \beta_k = 3.5$, $\beta_{k+1}, \ldots, \beta_{950} = 0$ for k equal 20, 100, 200;
- build models:
  - LASSO with
    - λ from CV,
    - λ from SURE;
  - Ridge Regression
    - λ from CV,
    - λ from SURE;
  - ElasticNet with $\alpha = 0.5$ and
    - λ from CV,
    - λ from SURE;
  - OLS
    - with all variables,
    - with variables selected by AIC,
    - with variables selected by mBIC2.
- evaluate the models by calculating MSE for $\hat{\beta}$ and $\hat{Y}$.

Simulations b

We will repeat the following experiment ten times:

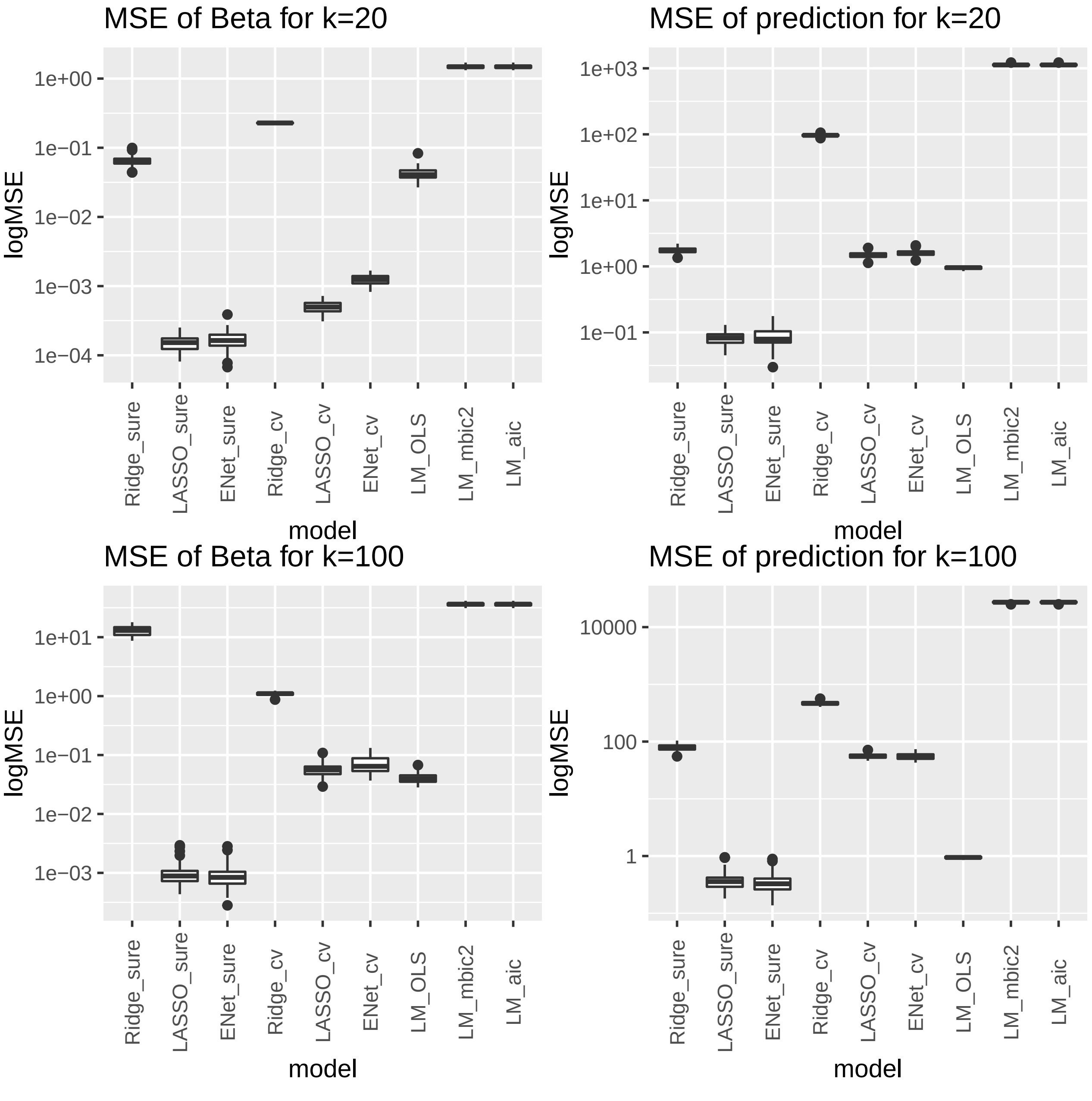
- sample matrix X of size 1000×950 from multivariate normal distribution $\mathcal{N}_n(0, \Sigma = 0.5 + 0.5 * \mathbb{1}i = 1)$, generate an error term vector from standard normal distribution;
- the real coefficients are: $\beta_1, \ldots, \beta_k = 5$, $\beta_{k+1}, \ldots, \beta_{950} = 0$ for k equal 20, 100, 200;
- build models:
  - LASSO with
    - λ from CV,
    - λ from SURE;
  - Ridge Regression
    - λ from CV,
    - λ from SURE;
  - ElasticNet with $\alpha = 0.5$ and
    - λ from CV,
    - λ from SURE;

- OLS
    - * with all variables,
    - * with variables selected by AIC,
    - * with variables selected by mBIC2.
- evaluate the models by calculating MSE for $\hat{\beta}$ and $\hat{Y}$.

In this and all next tasks, the AIC and mBIC2 criteria were used in the fast forward procedure (which may be suboptimal).

Please note that the number of iterations is lower than expected (100) because of the high time complexity of the task.

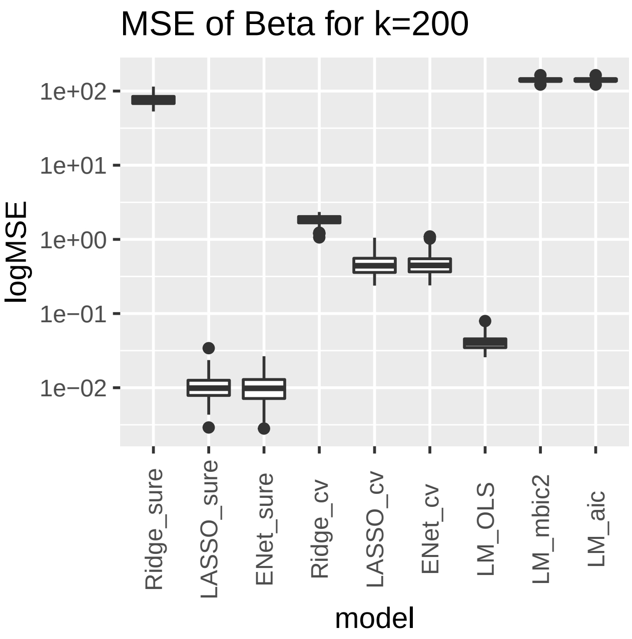
Results b


MSE of Beta for k=20

MSE of prediction for k=20

MSE of Beta for k=100

MSE of prediction for k=100



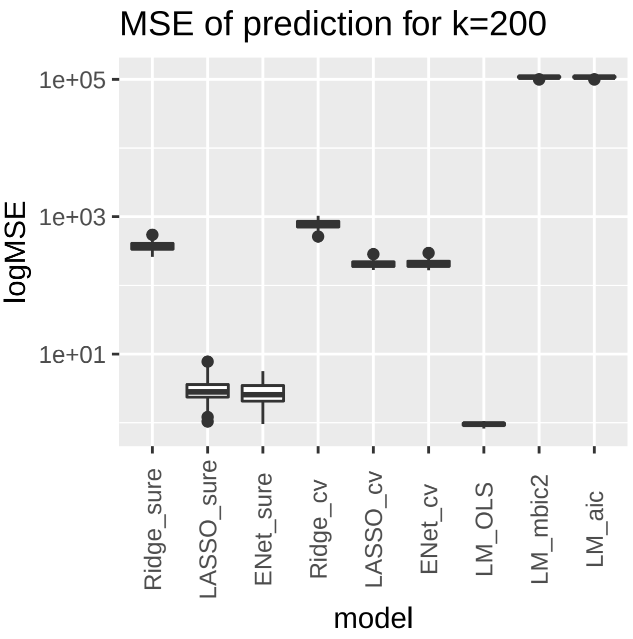


Table 6: MSE of Beta

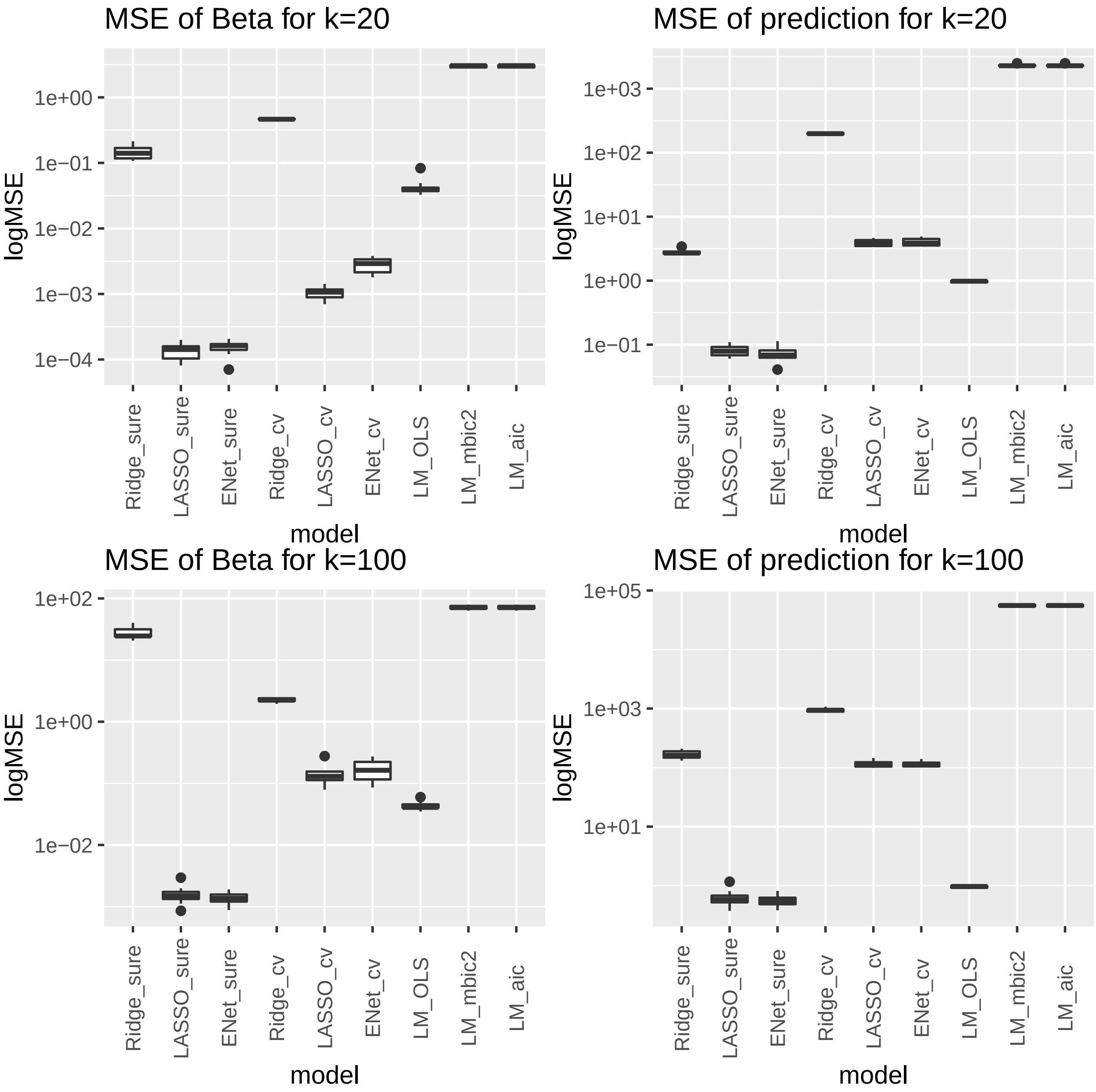
	20	100	200
Ridge_sure	0.06	13.11	76.26
LASSO_sure	0.00	0.00	0.01
ENet_sure	0.00	0.00	0.01
Ridge_cv	0.23	1.09	1.83
LASSO_cv	0.00	0.06	0.48
ENet_cv	0.00	0.07	0.49
LM_OLS	0.04	0.04	0.04
LM_mbic2	1.48	36.15	141.13
LM_aic	1.48	36.15	141.13

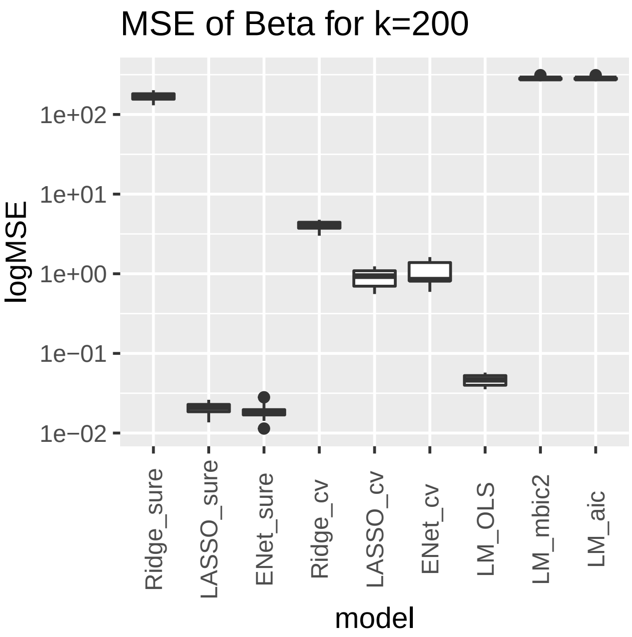
Table 7: MSE of prediction

	20	100	200
Ridge_sure	1.76	79.34	375.45
LASSO_sure	0.08	0.38	3.00
ENet_sure	0.09	0.35	2.80
Ridge_cv	96.48	466.78	784.16
LASSO_cv	1.49	55.65	207.95
ENet_cv	1.60	55.38	208.57
LM_OLS	0.95	0.94	0.95
LM_mbic2	1125.48	27233.39	108062.81
LM_aic	1125.48	27233.39	108062.81

When the variables are dependent, the task becomes much more complicated. We got satisfying results for methods performing L1 penalization (LASSO and ENet). Good predictions from OLS. Ridge performs poorly, OLS with preselected variables cannot handle this case.

Results b


MSE of Beta for k=20
logMSE
1e+00
1e−01
1e−02
1e−03
1e−04
Ridge_sure
LASSO_sure
ENet_sure
Ridge_cv
LASSO_cv
ENet_cv
LM_OLS
LM_mbic2
LM_aic
model
MSE of prediction for k=20
logMSE
1e+03
1e+02
1e+01
1e+00
1e−01
Ridge_sure
LASSO_sure
ENet_sure
Ridge_cv
LASSO_cv
ENet_cv
LM_OLS
LM_mbic2
LM_aic
model
MSE of Beta for k=100
logMSE
1e+02
1e+00
1e−02
Ridge_sure
LASSO_sure
ENet_sure
Ridge_cv
LASSO_cv
ENet_cv
LM_OLS
LM_mbic2
LM_aic
model
MSE of prediction for k=100
logMSE
1e+05
1e+03
1e+01
Ridge_sure
LASSO_sure
ENet_sure
Ridge_cv
LASSO_cv
ENet_cv
LM_OLS
LM_mbic2
LM_aic
model



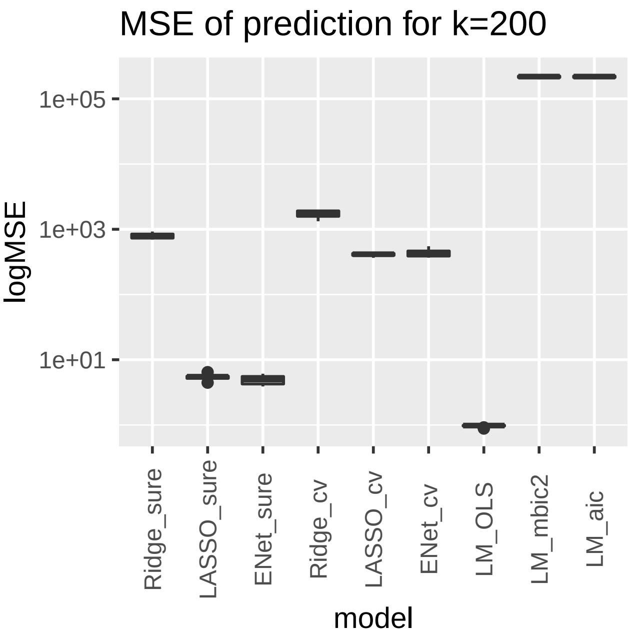


Table 8: MSE of Beta

	20	100	200
Ridge__sure	0.15	27.67	168.05
LASSO__sure	0.00	0.00	0.02
ENet__sure	0.00	0.00	0.02
Ridge__cv	0.47	2.26	3.98
LASSO__cv	0.00	0.14	0.91
ENet__cv	0.00	0.17	1.03
LM__OLS	0.04	0.04	0.05
LM__mbic2	3.01	71.82	284.38
LM__aic	3.01	71.82	284.38

Table 9: MSE of prediction

	20	100	200
Ridge__sure	2.78	167.80	794.44
LASSO__sure	0.08	0.63	5.43
ENet__sure	0.07	0.57	4.95
Ridge__cv	198.10	949.09	1708.51
LASSO__cv	3.91	116.25	408.13
ENet__cv	4.02	115.37	431.43
LM__OLS	0.97	0.96	0.97
LM__mbic2	2302.45	55565.47	216576.33
LM__aic	2302.45	55565.47	216576.33

For stronger signals, the observations from the previous example are further highlighted.

Task 5

In the fifth task we will ilustrate the irrepresentability nad identifiability properties.

Theoretical descritpion

Irrepresentability condition:

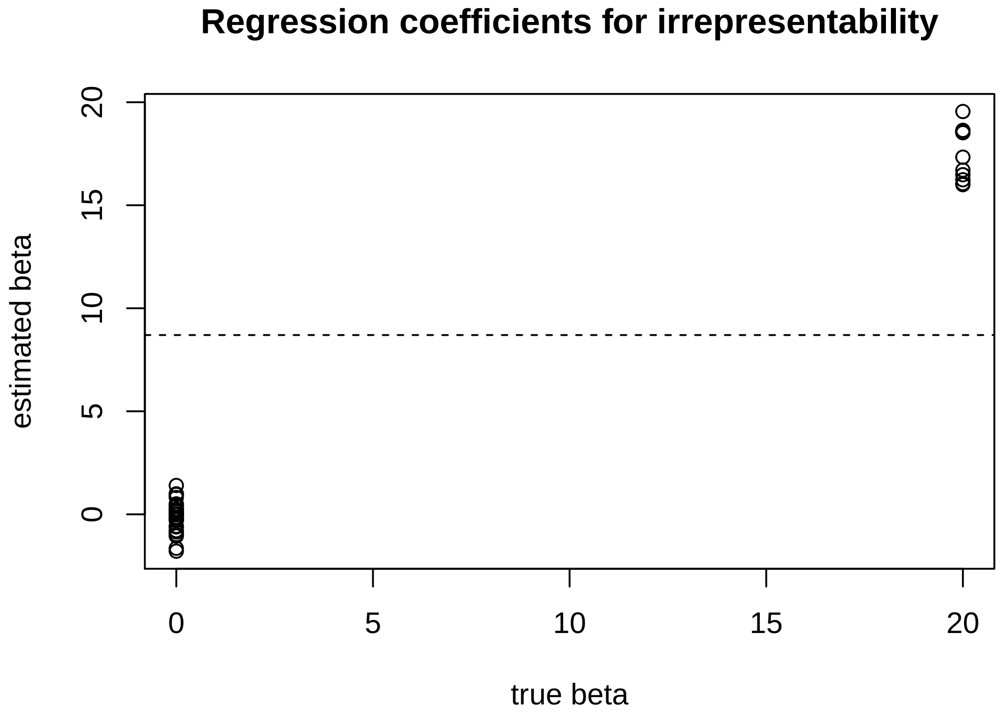
$$\|X_{\bar{I}}^T X_I (X_I^T X_I)^{-1} S_I\|_\infty \leq 1$$

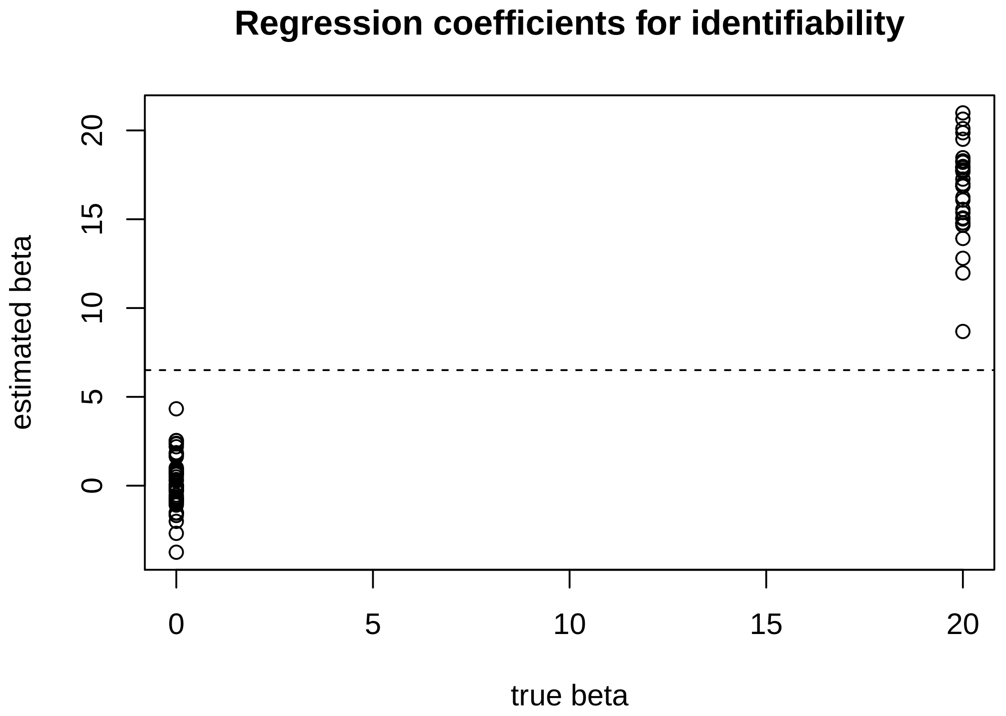
Identifiability condition:

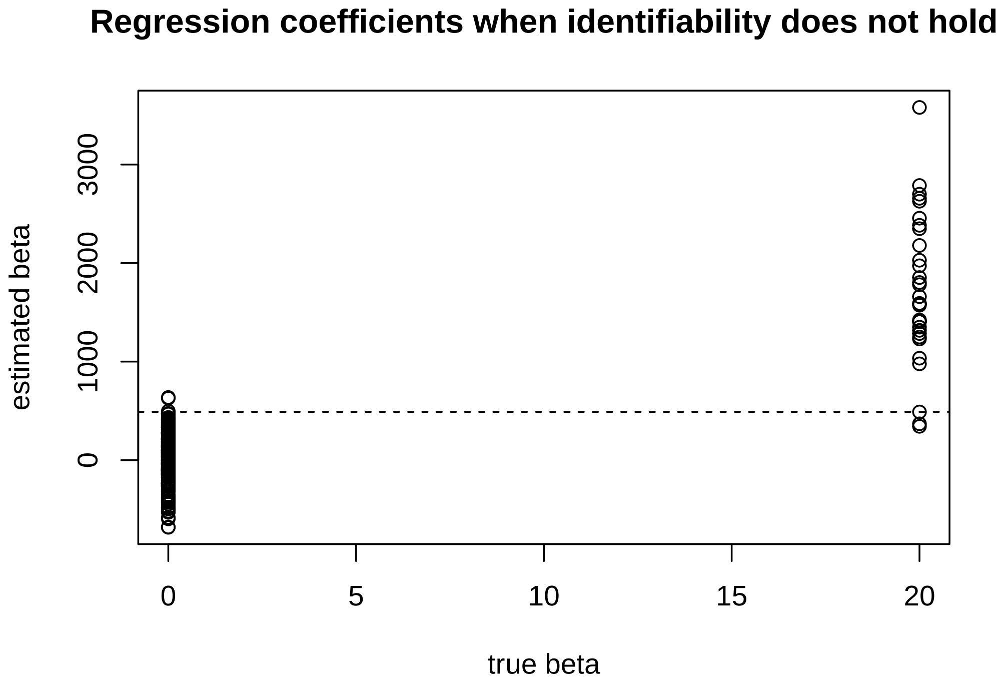
$$X\gamma = X\beta \text{ and } \gamma \neq \beta \text{ then } \|\gamma\|_1 > \|\beta\|_1$$

Simulation

- generate design matrix X fo size 100 × 200 form the normal distribution $\mathcal{N}(0, \sigma = 0.1)$;
- find the maximal k^{IR} for which the LASSO irrepresentability condition holds;
- find the maximal k^{ID} for which the LASSO identifiability condition holds;
- generate the vectors of coefficients:
  - $\beta_1, \ldots, \beta_k^{IR} = 20$, $\beta_{k+1}, \ldots, \beta_{200} = 0$,
  - $\beta_1, \ldots, \beta_k^{ID} = 20$, $\beta_{k+1}, \ldots, \beta_{200} = 0$,
  - $\beta_1, \ldots, \beta_k^{ID} = 20$, $\beta_{k+1}, \ldots, \beta_{200} = 0$;
- generate an error term from the standard normal distribution; for each of the above vectors, calculate the values of dependent variable Y;
- provide scatter plots for betas and their estimators.







As we can see in the above scatter plots, we can separate zero and nonzero elements of the beta vector when the irrepresentability and identifiability conditions hold. However, we cannot separate when the identifiability condition does not hold.

Task 6

In this task, we will apply techniques considered before to a real world example. We will predict the expression level of gene one using the expression levels of 3220 other genes.

Implementation

First we split the data into train (180 individuals) and test (30 individuals) set. Then we will build LASSO, RIdge and ENet models using lambda from cross-validation. First, we will build models using all variables, then we will build models using only preselected variables.

Results

Table 10: Evaluation of models build on all variables.

	Vars	RMSE	MAPE
Ridge	2304	0.2	1.13
LASSO	55	0.2	1.18
ENet	74	0.2	1.21

Table 11: Evaluation of models build on preselected variables.

	Vars	RMSE	MAPE
Ridge	295	0.2	1.12
LASSO	27	0.2	1.12
ENet	35	0.2	1.11

On both bases, Ridge Regression uses almost all available variables. It has similar results with and without preselecting variables. LASSO and ElasticNet perform better after preselecting variables; LASSO selects the lowest number od variables, ENet has the lowest Mean Absolute Percentage Error (MAPE).