Recent papers and collaborations

Collaboration Tomáš Řiháček, Masaryk University, Brno, CZ

- Řiháček et al. (in press). Mechanisms of change in multicomponent group-based treatment for patients suffering from medically unexplained physical symptoms. *Psychotherapy Research*. https://doi.org/10.1080/10503307.2022.2061874
- Pourová et al. (in press). Negative effects during multicomponent group-based treatment: A multisite study. *Psychotherapy Research*. https://doi.org/10.1080/10503307.2022.2095237

Central Institute of Mental Health, Mannheim, Ulrich Reininghaus, Public Mental Health:

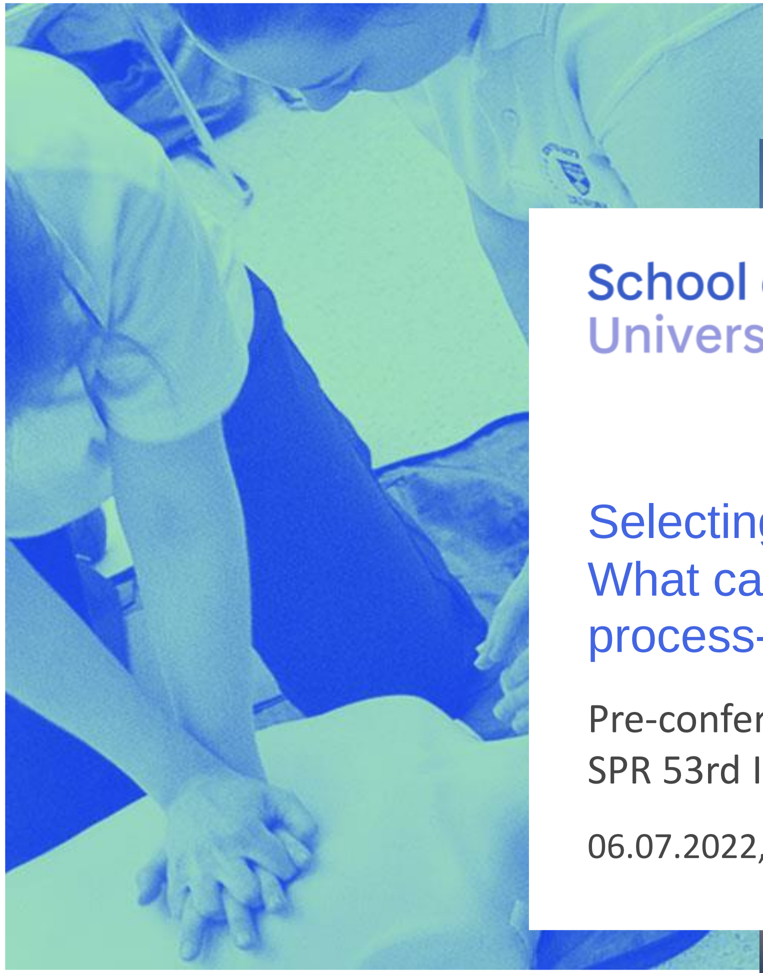
Schick et al. (2021). Effects of a novel, transdiagnostic, hybrid ecological momentary intervention for improving resilience in youth (EMIcompass). *JMIR Research Protocols, 10*(12). https://doi.org/10.2196/27462

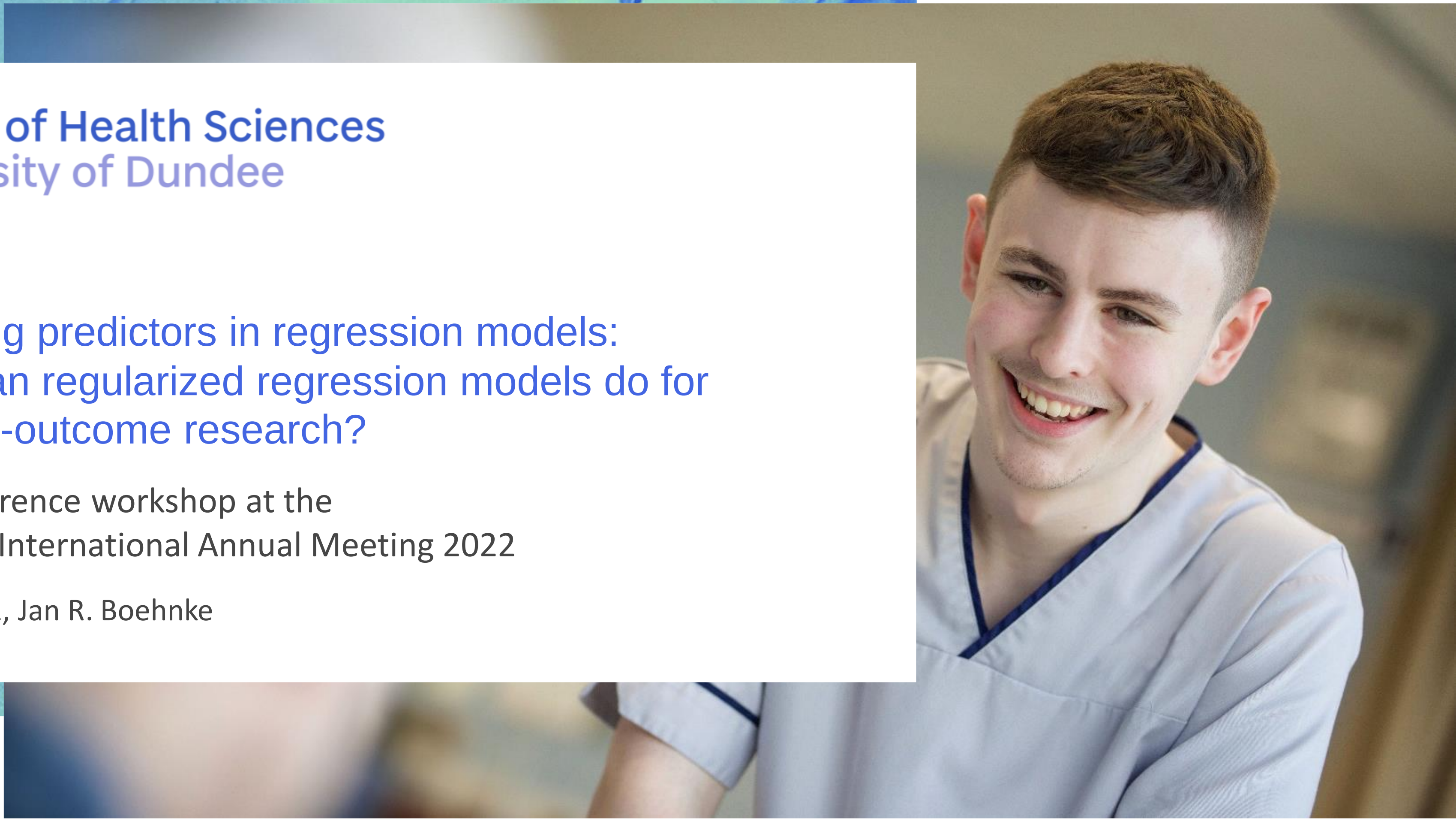
Anna Freud National Centre for Children and Families (Honorary Collaborator):

Mansfield et al. (2022). The impact of the COVID-19 pandemic on adolescent mental health: A natural experiment. *Royal Society Open Science*, 9(4). https://doi.org/10.1098/rsos.211114

Psychometric assessment and outcomes in mental health assessment:

Futures of health measurement: Core outcomes, item banks, and common measures, Schulich School of Medicine & Dentistry, Western University



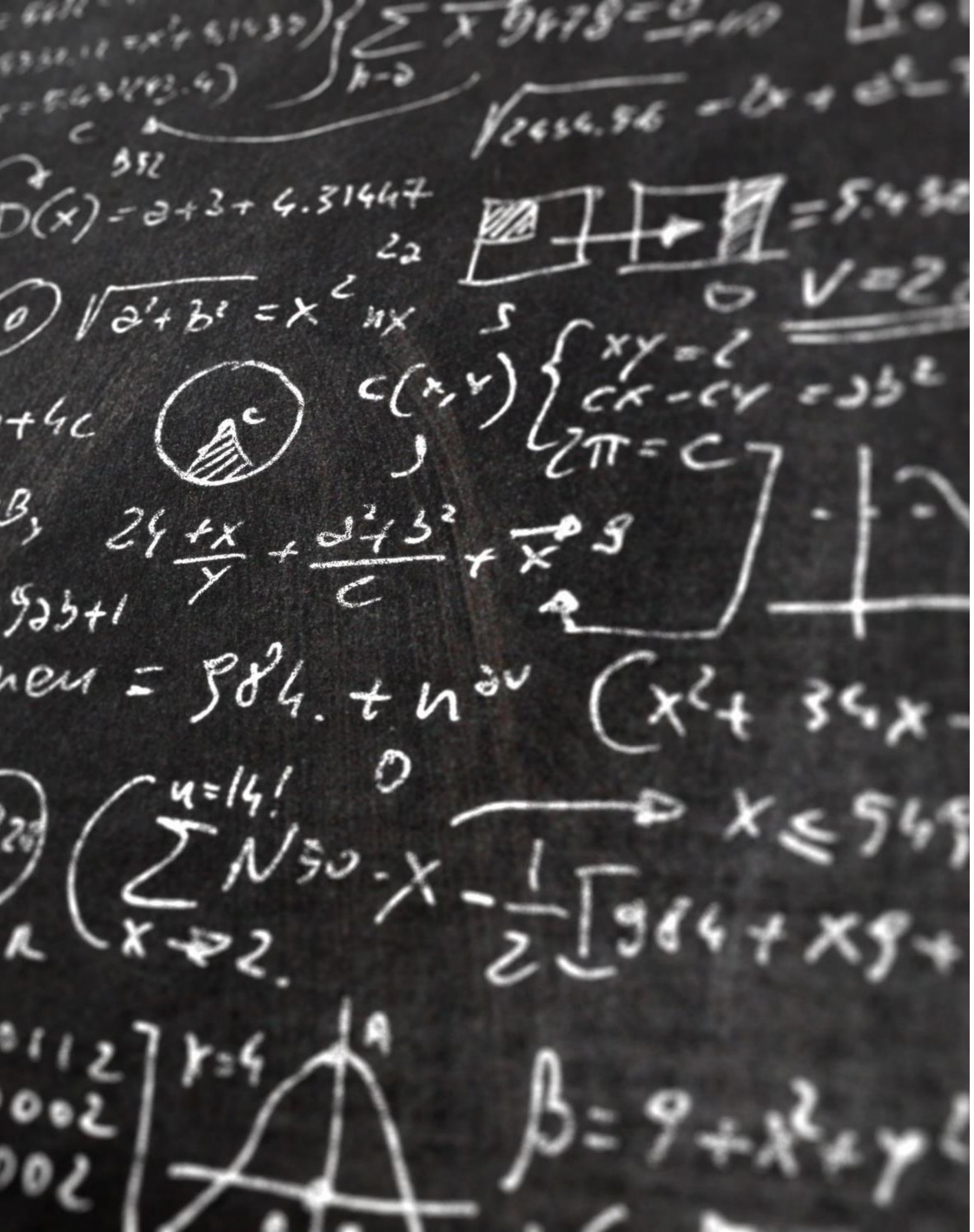


School of Health Sciences
University of Dundee

Selecting predictors in regression models: What can regularized regression models do for process-outcome research?

Pre-conference workshop at the
SPR 53rd International Annual Meeting 2022

06.07.2022, Jan R. Boehnke



Topics discussed

Differences between prediction and explanation

Briefly reviewing regression methods

Introduce basics of regularisation methods for feature selection

More general applications of regularization methods

Overview

Regression Analysis

- Brief re-cap
- OLS & selection
- Regularised models

Regularisation

- Lasso
- Elastic Net
- Ridge Regression
- Simulation Examples

Extensions & Discussion

- Many mediators
- SEM / CFA / IRT
- “Summary”

Photo by Mark König on Unsplash

Why perform a regression analysis?

“Multiple regression as a general data-analytic system.”

Cohen, 1968, Psychological Bulletin, 70, 426-443/ p. 426

“The linear regression model is the most commonly used statistical method in the social sciences.”

Long, 1997, *Regression models for categorical and limited dependent variables*. Sage, p. 1

“Regression analyses are a set of statistical techniques that allow one to assess the relationship between one DV [dependent variable] and several IVs [independent variables]. […] Regression techniques can be applied to a data set in which the IVs are correlated with one another and with the DV to varying degrees.”

Tabachnick & Fidell, 2007, *Using multivariate statistics* (5th edition). Boston: Allyn & Bacon. p. 117

Why perform a regression analysis?

$$\hat{Y} = b_0 + b_1 X + b_2 Z + e$$

The model assumes linear relation in its parameters.

The model assumes that there is a degree of linear independence between the independent variables.

The expected average error is “0”.

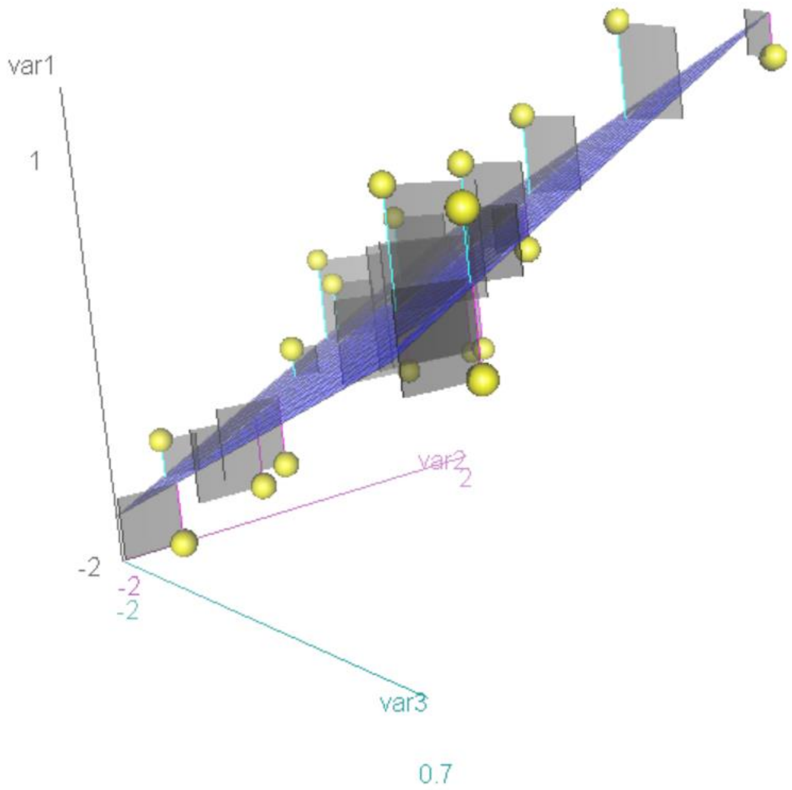
The conditional variance of the errors is constant.

Errors of individual observations are uncorrelated.

[The errors follow a normal distribution.]

Long, 1997, *Regression models for categorical and limited dependent variables*. Sage.

Why perform a regression analysis?



The model assumes linear relation in its parameters.

The model assumes that there is a degree of linear independence between the independent variables.

The expected average error is “0”.

The conditional variance of the errors is constant.

Errors of individual observations are uncorrelated.

[The errors follow a normal distribution.]

Long, 1997, *Regression models for categorical and limited dependent variables*. Sage.
Fox & Bouchet-Valat, 2021, Rcmdr: R Commander. R package version 2.7-2.

Prediction vs. explanation

Multiple regression can be used for exploratory or confirmatory purposes.

Multiple regression can be used for both predictive as well as explanatory purposes.

Since this has substantial effects on the appropriateness of applied methods, this needs to be stated clearly by the analysts.

Kelley & Maxwell, 2010, in G. R. Hancock & A. O. Mueller (Eds.), *The reviewer's guide to quantitative methods in the social sciences (pp. 281-297). Routledge.*

Prediction vs. Explanation

"*Description* is using data to provide a quantitative summary of certain features of the world."

"*Prediction* is using data to map some features of the world (the inputs) to other features of the world (the outputs)."

"*Counterfactual prediction* is using data to predict certain features of the world as if the world had been different, which is required in *causal inference* applications."

Hernán et al., 2019, *Chance, 32, 42-49. (all p. 43)*

Five primary ways in which GLMs for prediction differ from GLMs for causal inference:

(i) the covariates that should be considered for inclusion/exclusion;

(ii) the way how to identify a suitable set of covariates to include in the model;

(iii) which covariates are ultimately selected and what functional form they take;

(iv) how the model is evaluated; and

(v) how the model is interpreted.

Arnold et al., 2021, *International Journal of Epidemiology, 49, 2074-2082.*

Prediction vs. explanation

The utility of a **prediction** model lies in its ability to accurately predict the outcome of interest.

Such information may be used to anticipate the outcome

- to prepare for its occurrence/ estimated amount of occurrence
- inform a subsequent intervention that attempts to alter it (after the outcome has occurred!).

Which clients in a psychotherapy setting/ trial/… are most (or least) likely show reliable improvement?

The goal of **causal explanation** is to estimate the true causal association between a particular variable and the outcome

Multiple regression models allow to remove other hypothesized associations that distort that relationship.

Such information may then be used to attempt to alter the outcome by altering the exposure

Does an exposure intervention increase the probability of reliable improvement?

(…to which degree does which amount of exposure…)

Arnold et al., 2021, *International Journal of Epidemiology, 49, 2074-2082.*
Kelley & Maxwell, 2010, in G. R. Hancock & A. O. Mueller (Eds.), *The reviewer's guide to quantitative methods in the social sciences (pp. 281-297). Routledge.*

Prediction vs. Explanation

Variables that are hypothesized to be **useful ‘predictors’** of the outcome should be identified; these are variables that are likely to be associated with the outcome, though not necessarily directly causally related to it.

Practical considerations

- variables that appear in a dataset are considered for inclusion;
- variables that are easy to measure and/or record;
- variables for which high measurement quality can be achieved (very broadly interpreted).

The **causal association of interest** is represented by the coefficient of the exposure variable; removing all spurious associations is achieved in principle by also including as covariates a sufficient set of variables that ‘control for’ those associations.

- Literature review and other work to justify selections;
- Directed acyclic graphs as a way to formalise the findings and/or assumptions.

Equally important to identify variables for inclusion as well as variables to exclude (e.g. blocking causal paths or colliders).

Arnold et al., 2021, *International Journal of Epidemiology, 49, 2074-2082.*

A good final set of predictors and their interpretation

Variables that are finally included in a **prediction model** are those that together efficiently maximize the amount of information relative to the outcome.

This ‘optimal’ subset of covariates usually offers a trade-off between ‘explaining variation’ in the outcome and being parsimonious enough so that is likely to fit similar datasets.

Interpretation: The prediction model provides information about the expected value (or risk) of an outcome, given data on the covariates in the model.

The model does not provide information about how to change the expected value (or risk) of an outcome.

Arnold et al., 2021, *International Journal of Epidemiology, 49, 2074-2082.*

Prediction vs. explanation

Workshop today far into "prediction" space…

Literature

Regression Models

- Cohen, J., Cohen, P., West, S.G., & Aiken, L.S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd edition). Mahwah, NJ: Lawrence Erlbaum.
- Fox, J., & Weisberg, S. (2011). *An R companion to applied regression* (2nd edition ed.). Sage.
- Long, J.S. (1997). *Regression models for categorical and limited dependent variables*. Thousand Oaks, CA: Sage.
- Tabachnick, B.G. & Fidell, L.S. (2007). *Using multivariate statistics* (5th edition). Boston: Allyn & Bacon.

Prediction vs. explanation

- Arnold, K. F., Davies, V., de Kamps, M., Tennant, P. W. G., Mbotwa, J., & Gilthorpe, M. S. (2021). Reflection on modern methods: generalized linear models for prognosis and intervention-theory, practice and implications for machine learning. *Int J Epidemiol, 49*, 2074-2082.
- Tennant, P. W. G., Murray, E. J., Arnold, K. F., Berrie, L., Fox, M. P., Gadd, S. C., Harrison, W. J., Keeble, C., Ranker, L. R., Textor, J., Tomova, G. D., Gilthorpe, M. S., & Ellison, G. T. H. (2021). Use of directed acyclic graphs (DAGs) to identify confounders in applied health research: review and recommendations. *Int J Epidemiol, 50*, 620-632.
- Rohrer, J. M. (2018). Thinking Clearly About Correlations and Causation: Graphical Causal Models for Observational Data. *Advances in Methods and Practices in Psychological Science, 1, 27-42.*
- Wysocki, A. C., Lawson, K. M., & Rhemtulla, M. (2022). Statistical Control Requires Causal Justification. *Advances in Methods and Practices in Psychological Science, 5(2).* https://doi.org/10.1177/25152459221095823

Data set

Tab-separated text file with simulated data set “SPR_Teaching_data.txt”

- N=250
- 15 variables, first one (var1) used as the dependent variable
- offers statistical power to test individual coefficients down to r=0.20 at power of beta=.90

And there is a cross-validation data set we are using:

“SPR_Teaching_CVdata.txt”

Materials will be available on https://github.com/pschikkolog/SPR_regularised/

Slides available via ResearchGate.

Ordinary least squares regression

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
Intercept  0.006377   0.052441   0.122   0.9033
var2       0.406720   0.055409   7.340 3.46e-12 ***
var3       0.268028   0.055050   4.869 2.06e-06 ***
var4       0.136309   0.057186   2.384   0.0179 *
var5      -0.007961   0.056565  -0.141   0.8882
var6      -0.066250   0.053887  -1.229   0.2201
var7      -0.019385   0.054748  -0.354   0.7236
var8       0.047422   0.053617   0.884   0.3774
var9      -0.058619   0.058277  -1.006   0.3155
var10      0.013977   0.052264   0.267   0.7894
var11      0.018527   0.057758   0.321   0.7487
var12     -0.034677   0.054923  -0.631   0.5284
var13      0.091417   0.056246   1.625   0.1054
var14      0.076580   0.059878   1.279   0.2022
var15      0.028761   0.055776   0.516   0.6066
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Residual standard error = 0.7869

with 235 degrees of freedom

Multiple R-squared: 0.4317

Adjusted R-squared: 0.3979

F-statistic: 12.75 (df1=14, df2=235), $p < 0.001$

Univariate screening

```
> p.table
   variable p-value
1      var2   0.000
2      var3   0.000
3      var4   0.000
4      var5   0.010
5      var6   0.005
6      var7   0.004
7      var8   0.003
8      var9   0.000
9     var10   0.000
10    var11   0.000
11    var12   0.000
12    var13   0.000
13    var14   0.000
14    var15   0.000
> |
```

Descriptively, bivariate relationships are interesting and can serve a lot of purposes, but…

Problems:

- multiplicity / multiple test
- presence of mediation effects
- presence of suppression effects
- …

Backward selection

```
 Deleted Chi-Sq d.f. P      Residual d.f. P      AIC    R2
 var5    0.02   1    0.8881 0.02      1   0.8881  -1.98 0.432
 var10   0.07   1    0.7983 0.09      2   0.9583  -3.91 0.432
 var11   0.12   1    0.7276 0.21      3   0.9765  -5.79 0.431
 var7    0.10   1    0.7500 0.31      4   0.9893  -7.69 0.431
 var15   0.23   1    0.6345 0.53      5   0.9908  -9.47 0.430
 var12   0.38   1    0.5371 0.91      6   0.9886 -11.09 0.430
 var8    0.76   1    0.3822 1.68      7   0.9755 -12.32 0.428
 var9    0.78   1    0.3766 2.46      8   0.9635 -13.54 0.426
 var6    1.22   1    0.2691 3.68      9   0.9311 -14.32 0.423
 var14   1.35   1    0.2444 5.04     10   0.8887 -14.96 0.420
 var13   3.48   1    0.0621 8.52     11   0.6663 -13.48 0.411

Approximate Estimates after Deleting Factors

               Coef    S.E.  Wald Z         P
Intercept -0.008705 0.05055 -0.1722 8.633e-01
var2       0.427349 0.05076  8.4191 0.000e+00
var3       0.265138 0.05039  5.2617 1.428e-07
var4       0.147683 0.05382  2.7440 6.070e-03

Factors in Final Model

[1] var2 var3 var4
```

Selection process here:

Out-selection of individual coefficients from full model based on $p < 0.05$ for the Wald-test of each individual variable.

Regularisation:
Enter the “lasso”

LASSO regression
(least absolute shrinkage and selection operator)

Regularization methods provide a means to constrain (“regularize”) the estimated coefficients, which can reduce the variance and decrease out-of-sample error.

Standard OLS regression

$$\text{minimize}\left(SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2\right)$$

Ridge regression

$$\text{minimize}\left(SSE + \lambda\sum_{j=1}^{p}\beta_j^2\right)$$

Lasso regression

$$\text{minimize}\left(SSE + \lambda\sum_{j=1}^{p}|\beta_j|\right)$$

Elastic net regression

$$\text{minimize}\left(SSE + \lambda_1\sum_{j=1}^{p}\beta_j^2 + \lambda_2\sum_{j=1}^{p}|\beta_j|\right)$$

Suggested reads regarding forms and implementation:

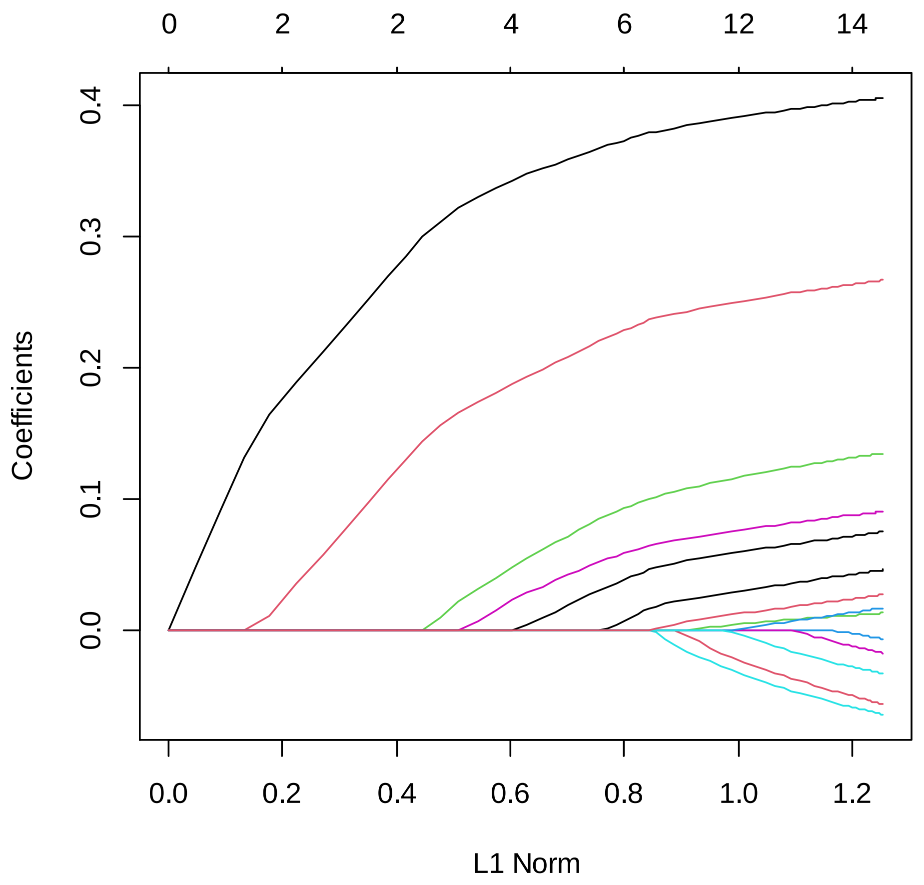
Boehmke & Greenwell, 2020, Hands-On Machine Learning with R. https://bradleyboehmke.github.io/HOML/index.html
(Formal representation taken from their Chapter 6)

Hastie et al., 2021, An Introduction to glmnet: https://glmnet.stanford.edu/articles/glmnet.html

LASSO regression
(least absolute shrinkage and selection operator)

```
lasso.example <-
glmnet(y=as.vector(teach.dat[ , 1]),
x=as.matrix(teach.dat[ , 2:15]))

plot(lasso.example)
```



This plot starts on the left hand side with all predictors weighted at "0".

Each line deviating from the horizontal "0"-line is a variable that is added into the model at a specific value of the L1-Norm ("lambda").

Each line presents the value of the regression coefficient for the variable the L1-norm increases (x-axis)

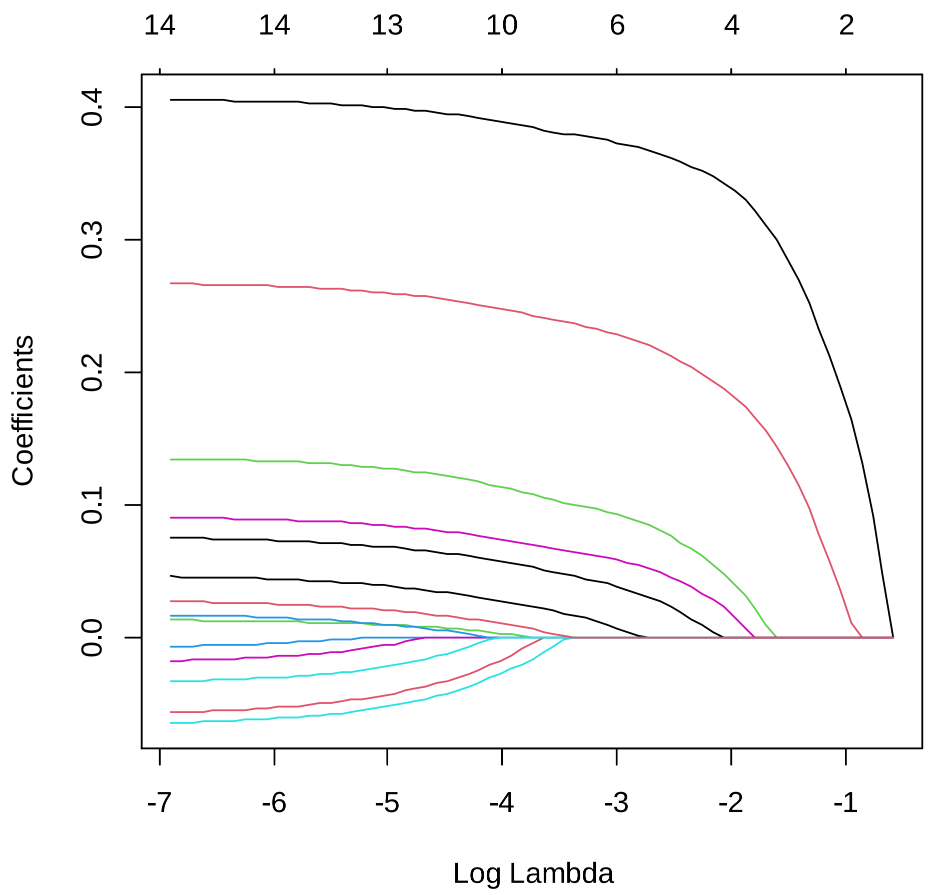
Inspecting result of lasso regression

```
#Plot lamda on log scale:
plot(lasso.example, "lambda")
```

Presentation on log-lambda scale.

A very nice explanation can also be found here:

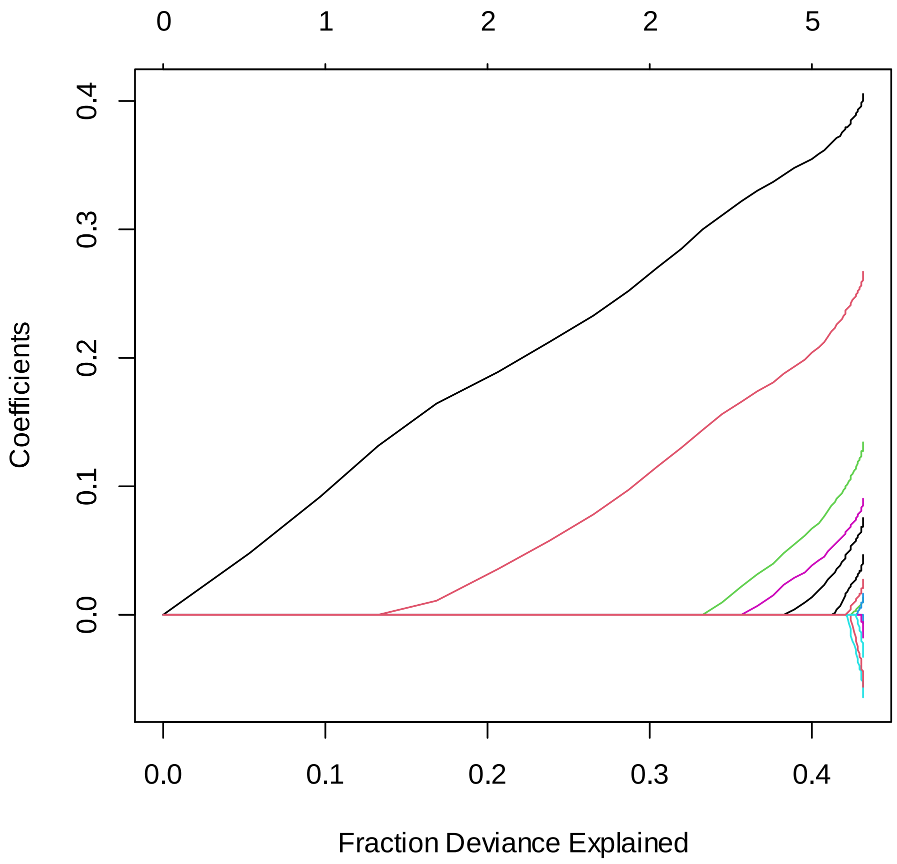
https://stats.stackexchange.com/questions/68431/interpretting-lasso-variable-trace-plots



Inspecting result of lasso regression

Another way of looking at the results obtained so far is to retrieve the table that provides the deviance measure depending on lambda/ the L1-norm

```
plot(lasso.example,
"dev")
```



	Df	%Dev	Lambda
1	0	0.00	0.56360
2	1	5.27	0.51360
3	1	9.64	0.46790
4	1	13.27	0.42640
5	2	16.88	0.38850
6	2	20.68	0.35400
7	2	23.84	0.32250
8	2	26.47	0.29390
9	2	28.64	0.26780
10	2	30.45	0.24400
11	2	31.95	0.22230
12	2	33.20	0.20260
13	3	34.46	0.18460
14	3	35.59	0.16820
15	4	36.65	0.15320
16	4	37.55	0.13960
17	4	38.30	0.12720
18	5	38.95	0.11590
19	5	39.52	0.10560
20	5	39.99	0.09623
21	5	40.38	0.08769
22	5	40.70	0.07990
23	5	40.97	0.07280
24	5	41.19	0.06633
25	6	41.39	0.06044
26	6	41.56	0.05507
27	6	41.70	0.05018
28	6	41.82	0.04572

Inspecting result of lasso regression

We can the use the coef-method to investigate which variables are included (and with which regression coefficients) if we choose a specific lambda-value

For example at L1=.368 the model would only include:

- the intercept (-.07) and
- var2 with a coeff of 0.18
- and var 3 with coeff = 0.03

HOW DO WE CHOOSE LAMDA/L1?

```
> coef(lasso.example, s=exp(-1))
15 x 1 sparse Matrix of class "dgCMatrix"
                     s1
(Intercept) -0.07244933
var2         0.17941295
var3         0.02545869
var4         .
var5         .
var6         .
var7         .
var8         .
var9         .
var10        .
var11        .
var12        .
var13        .
var14        .
var15        .
> coef(lasso.example, s=exp(-2))
15 x 1 sparse Matrix of class "dgCMatrix"
                     s1
(Intercept) -0.03197314
var2         0.33885346
var3         0.18318287
var4         0.04306529
var5         .
var6         .
var7         .
var8         .
var9         .
var10        .
var11        .
var12        .
var13        0.01803674
var14        .
```

How to choose L1/lamda?

The package includes an automated cross-validation procedure.

The function runs glmnet nfolds+1 (see next slide for nfolds) times

- the first run is to get the lambda sequence
- the remainder are used to compute the fit with each of the folds omitted
- the prediction error of the selected model error is accumulated, and the average error and standard deviation over the folds is computed for each of the runs.

```
lasso.cv.example <-
cv.glmnet(y=as.vector(teach.dat[ , 1]),
x=as.matrix(teach.dat[ , 2:15]),
nfold=10)
```

```
plot(lasso.cv.example)
```

#Instead of "nfold=" with a number also a variable providing groups could be specified; use "foldid=" instead

```
Measure: Mean-Squared Error

     Lambda Index Measure      SE Nonzero
min 0.04572    28  0.6235 0.05531       6
lse 0.16817    14  0.6784 0.06614       3
```

How to choose L1/lamda?

At the top the figure shows how many variables are included as L1 increases (see log scale on bottom!)

Two criteria typically used to determine the optimal L1:

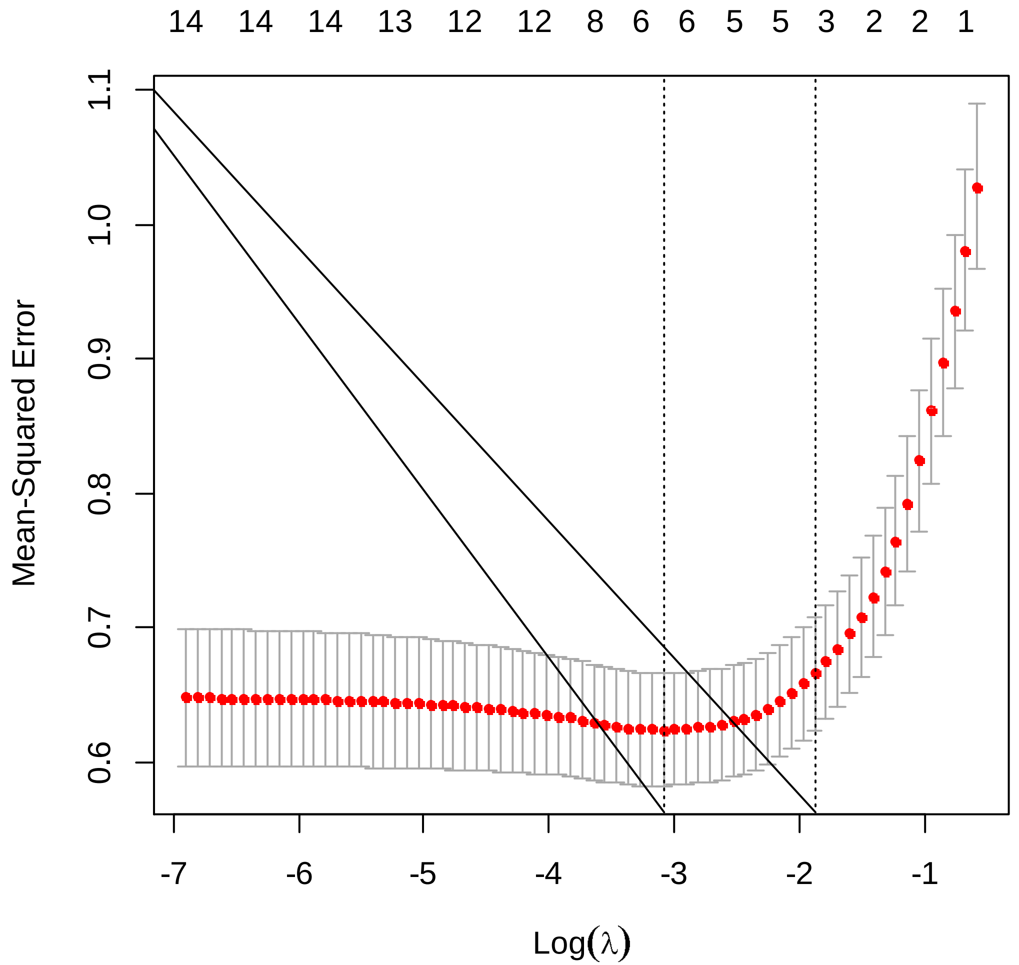
• lamda.min is the L1 that results in the smallest average cross-validation error in the cross-validation samples

• lambda+1SE is the lamda that is 1standard error larger than lambda.min

```
#These are plotted into the figure (vertical lines) and can be obtained via:

log(cv.lasso.1$lambda.min)

log(cv.lasso.1$lambda.1se)
```



The resulting model according to lasso regression

```
coef(lasso.cv.example, s =
"lambda.1se")
```

Selecting the model at 1SE above the lowest prediction error.

```
> coef(lasso.cv.example, s = "lambda.1se")
15 x 1 sparse Matrix of class "dgCMatrix"
                     s1
(Intercept) -0.04730273
var2         0.29983813
var3         0.14430940
var4        .
var5        .
var6        .
var7        .
var8        .
var9        .
var10       .
var11       .
var12       .
var13       .
var14       .
var15       .
> 
```

Results: So what is the real structure?

Given our sample and predictors as well as the cross-validation procedure chosen, lasso regression suggests that six predictors are the most relevant ones.

"Most relevant" in this case means those predictors for whom the loss in predictive value in the cross-validation sample is within a small margin (1 SE) compared to the model resulting in the smallest possible average prediction error.

Variables that are finally included in a **prediction model** are those that together efficiently maximize the amount of information relative to the outcome.

This ‘optimal’ subset of covariates usually offers a trade-off between ‘explaining variation’ in the outcome and being parsimonious enough so that is likely to fit similar datasets.

Arnold et al., 2021, *International Journal of Epidemiology, 49, 2074-2082.*

Results: So what is the real structure?

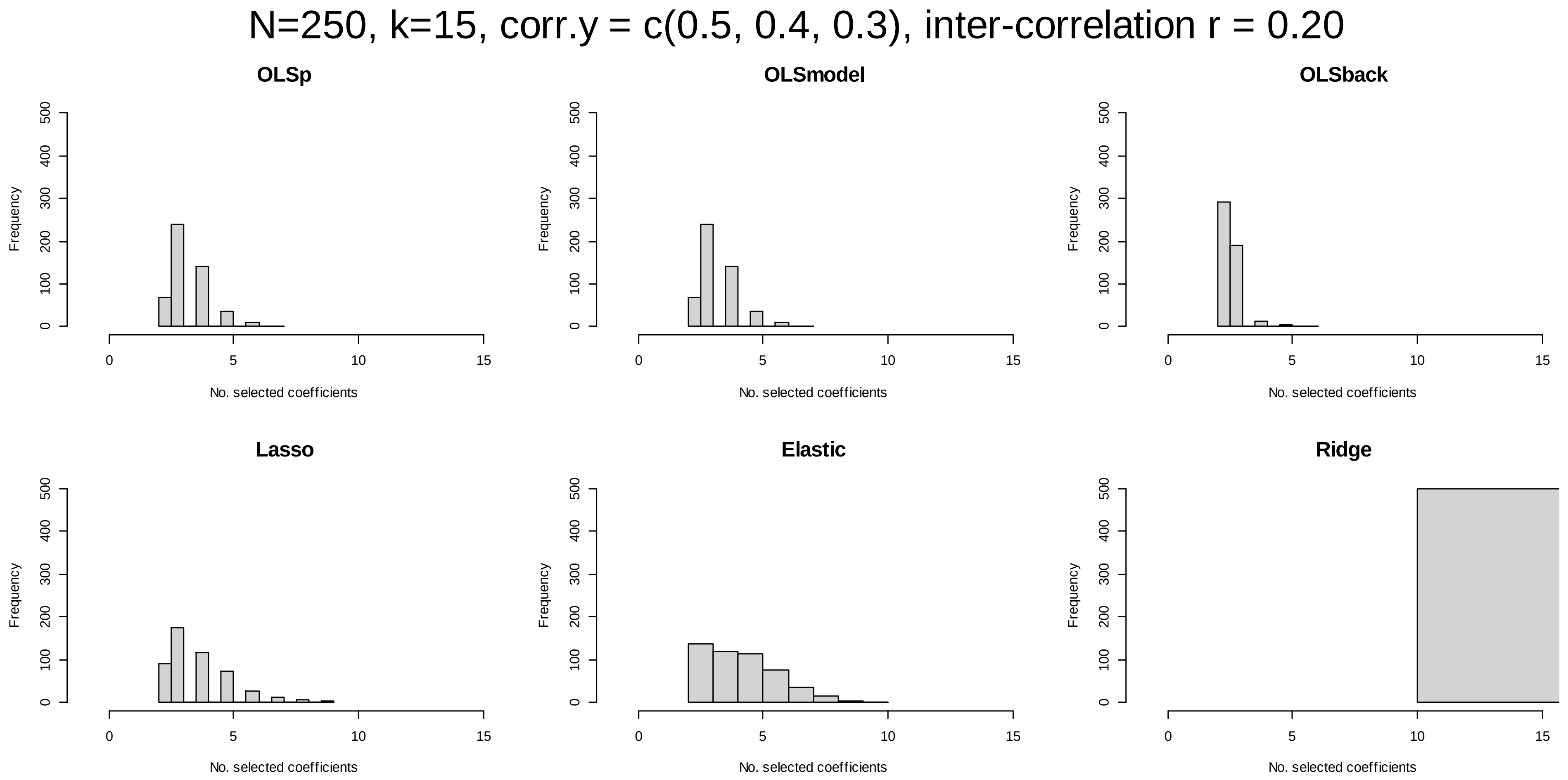
Variable	OLS	Backward	Lasso (α=1)	Elastic net (α=0.5)	Ridge (α=0)
var2	**0.41**	0.43	0.30	0.29	0.16
var3	**0.27**	0.27	0.14	0.16	0.12
var4	**0.14**	0.15	-	0.04	0.07
var5	-0.01	-	-	-	0.01
var6	-0.07	-	-	-	0.01
var7	-0.02	-	-	-	0.02
var8	0.05	-	-	-	0.03
var9	-0.06	-	-	-	0.02
var10	0.01	-	-	-	0.03
var11	0.02	-	-	-	0.03
var12	-0.03	-	-	-	0.04
var13	0.09	-	-	0.02	0.06
var14	0.08	-	-	-	0.05
var15	0.03	-	-	-	0.04

Results: So what is the real structure?

Variable	OLS	Backward	Lasso (α=1)	Elastic net (α=0.5)	Ridge (α=0)
Correlation	0.53	NA	0.55	0.56	0.50
Mean squared error	0.69	NA	0.70	0.69	0.73

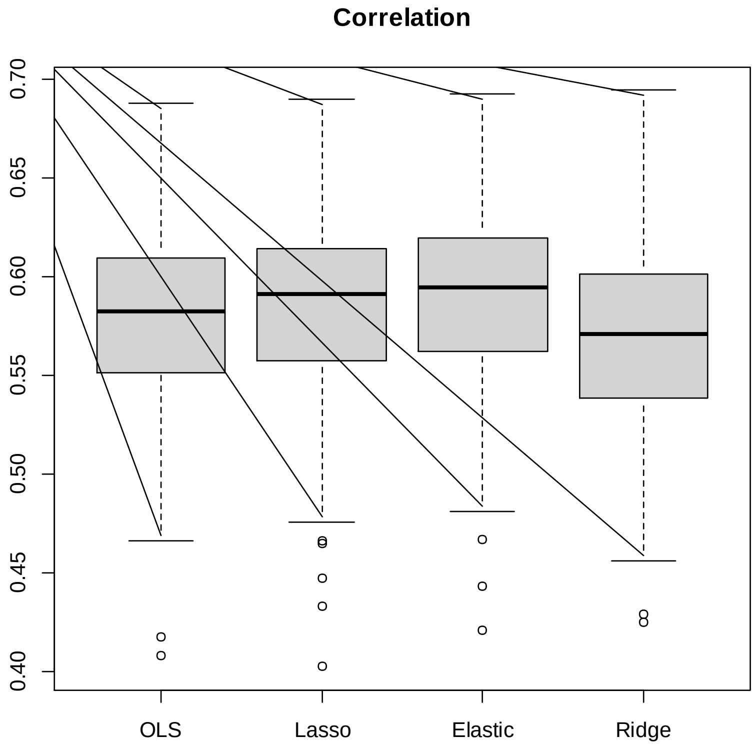
Let’s look at some simulations

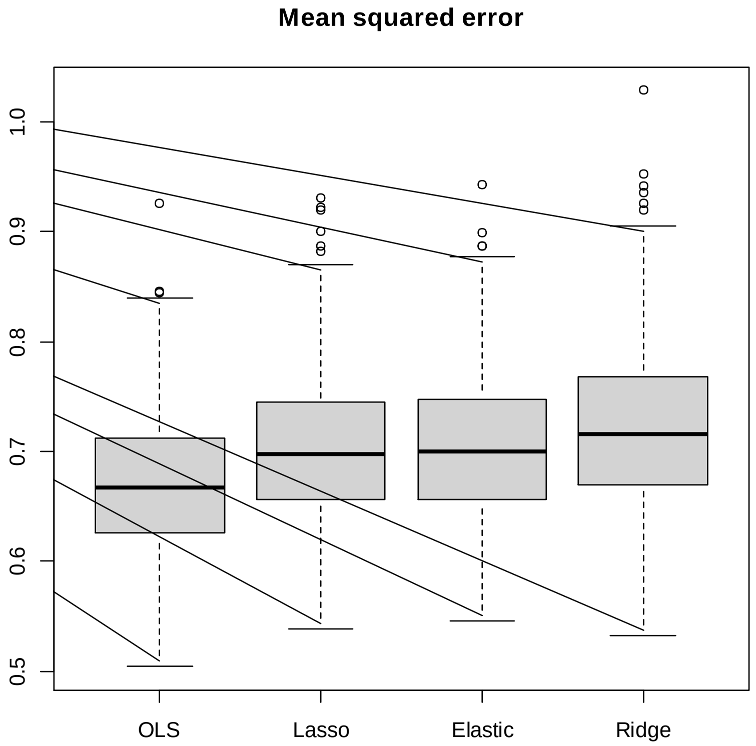
Same structure as the teaching example



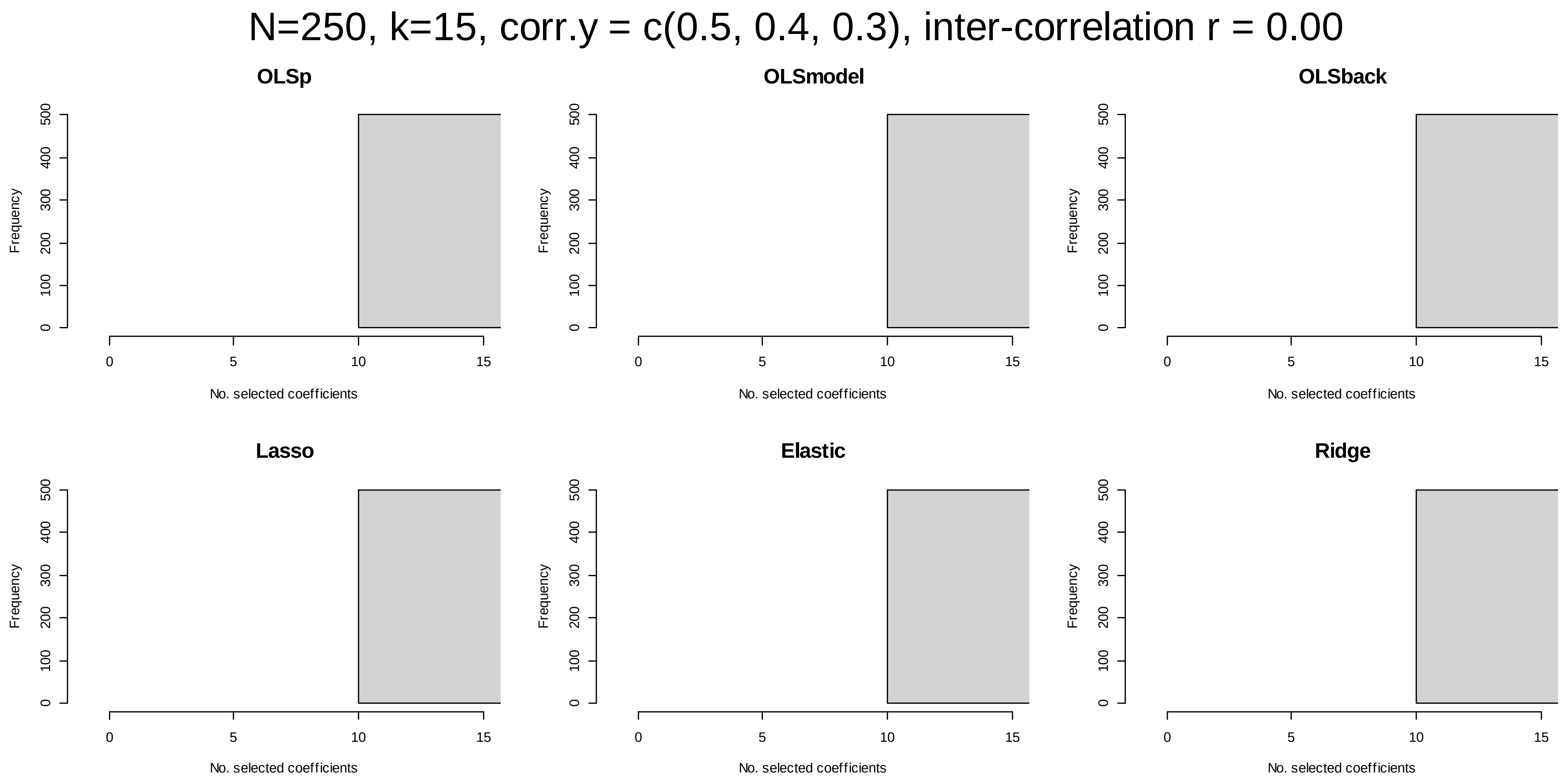
Same structure as the teaching example

```
   OLSp            OLSback         Lasso           Elastic         Ridge
 Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode:logical
 FALSE:500       FALSE:500       FALSE:500       FALSE:500       TRUE:500
```



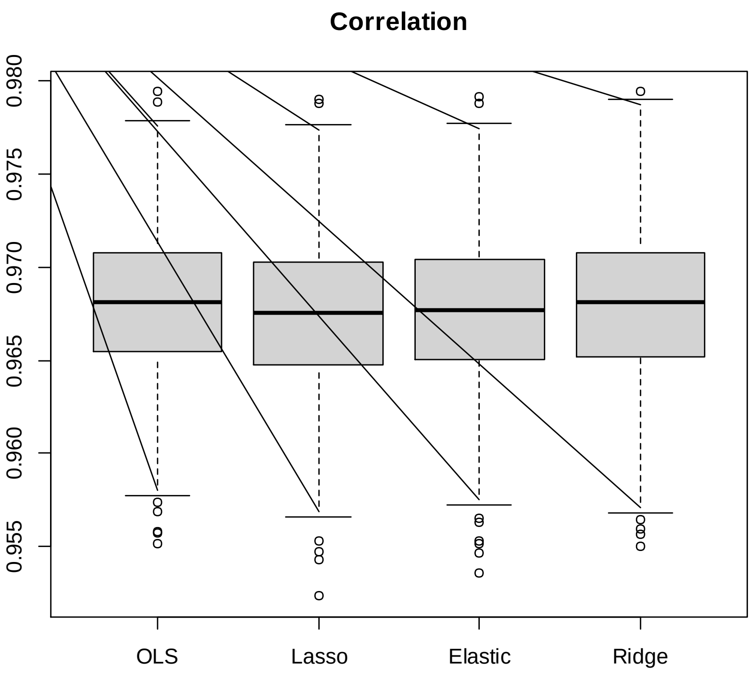


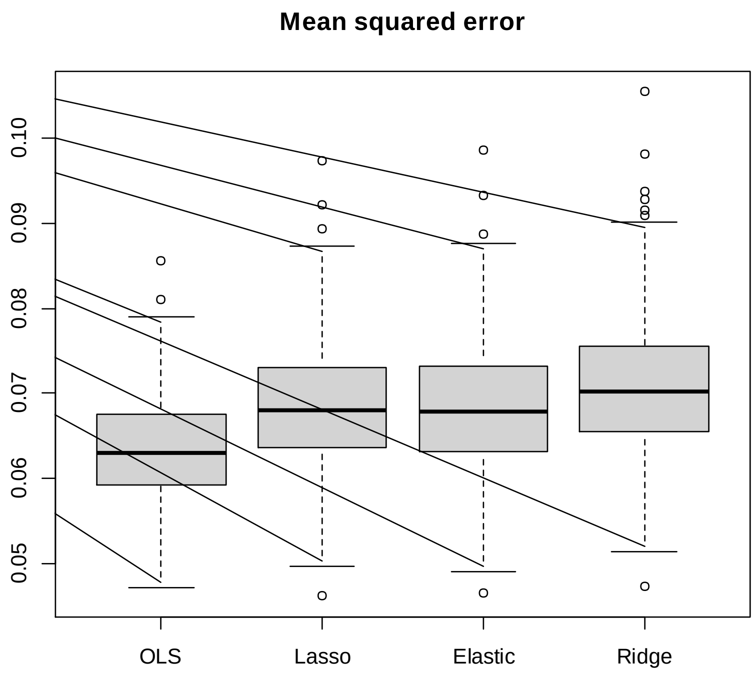
Teaching example, but no inter-correlations between X



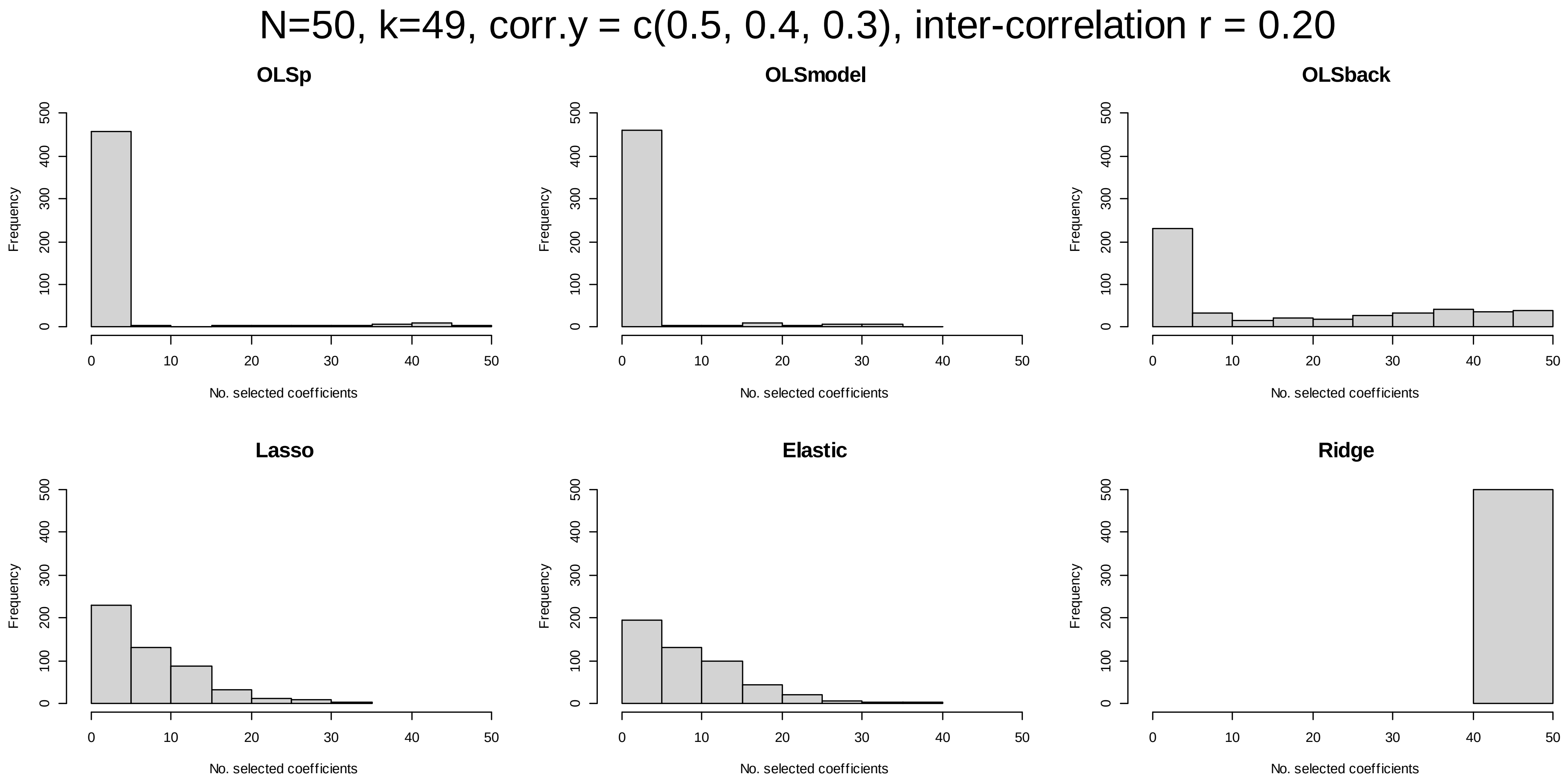
Teaching example, but no inter-correlations between X

```
  OLSp          OLSback        Lasso          Elastic        Ridge
Mode:logical   Mode:logical   Mode:logical   Mode:logical   Mode:logical
TRUE:500       TRUE:500       TRUE:500       TRUE:500       TRUE:500
```



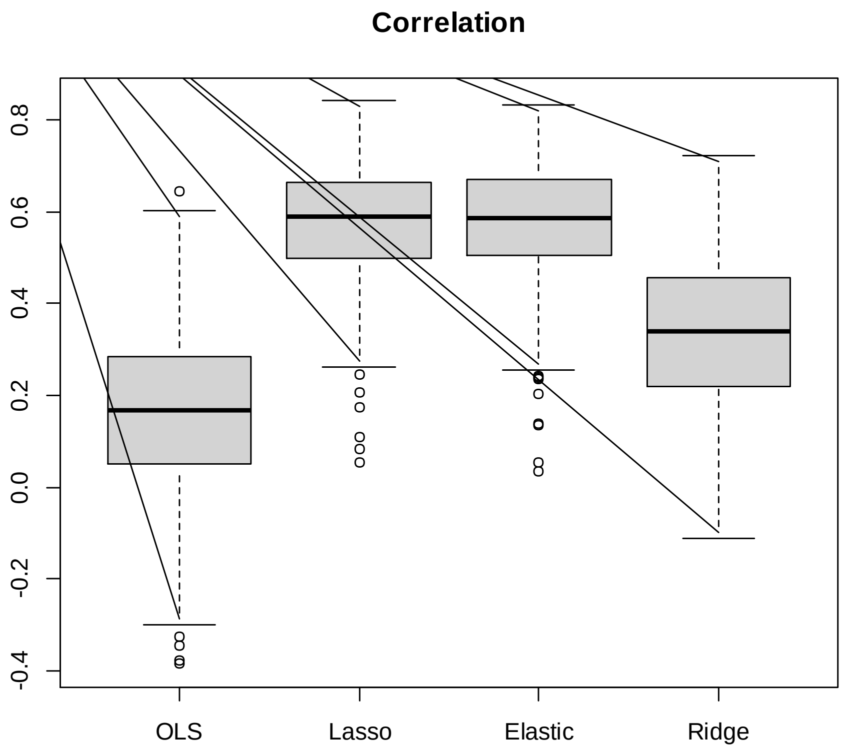


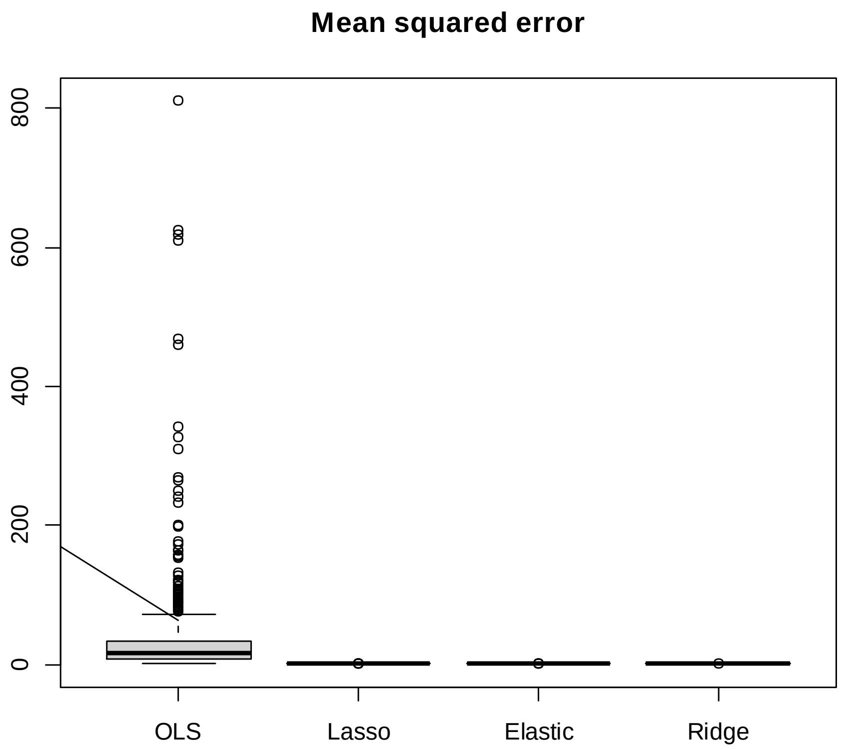
Small N, high p, Case 1



Small N, high p, Case 1

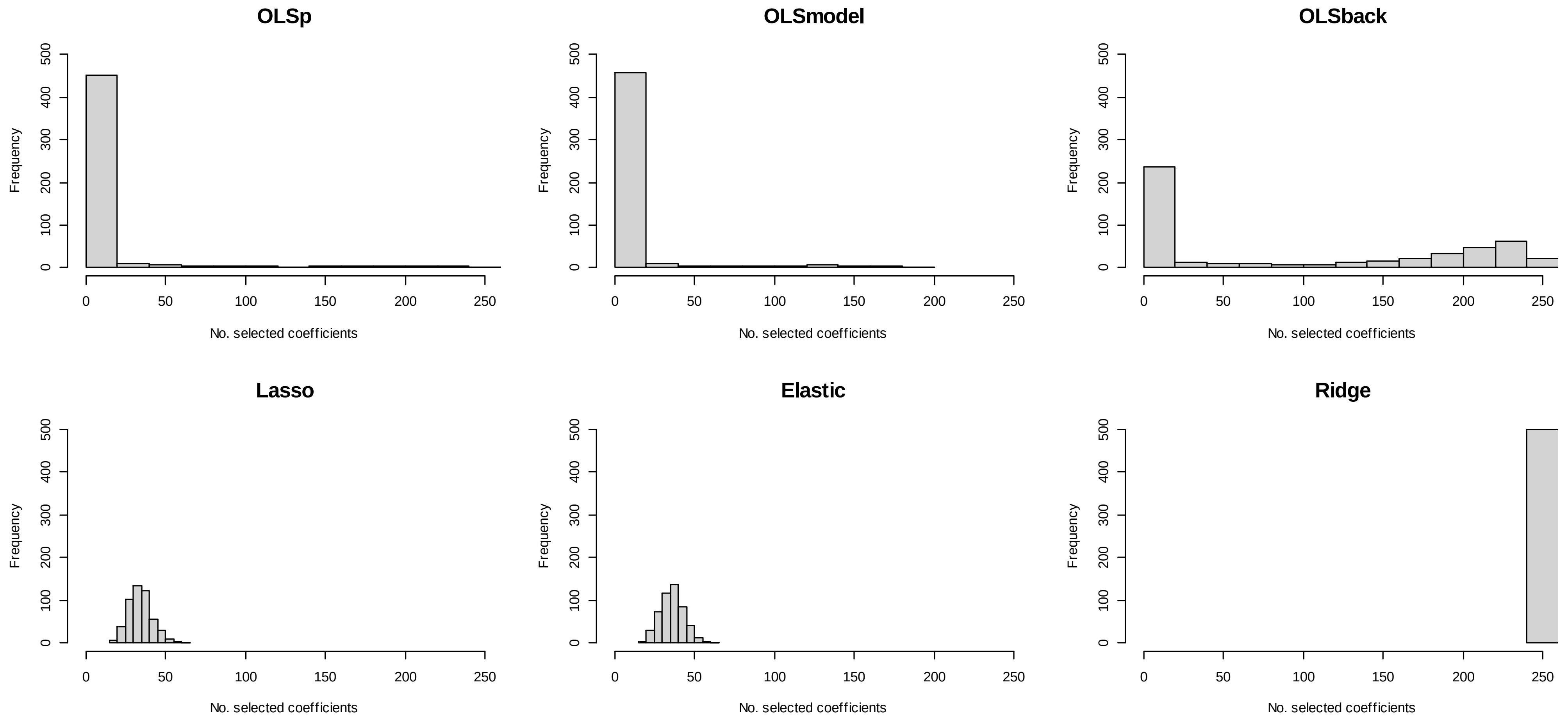
```
   OLSp              OLSback           Lasso             Elastic           Ridge
 Mode  :logical    Mode  :logical    Mode  :logical    Mode  :logical    Mode  :logical
 FALSE:499         FALSE:500         FALSE:487         FALSE:479         FALSE:500
```





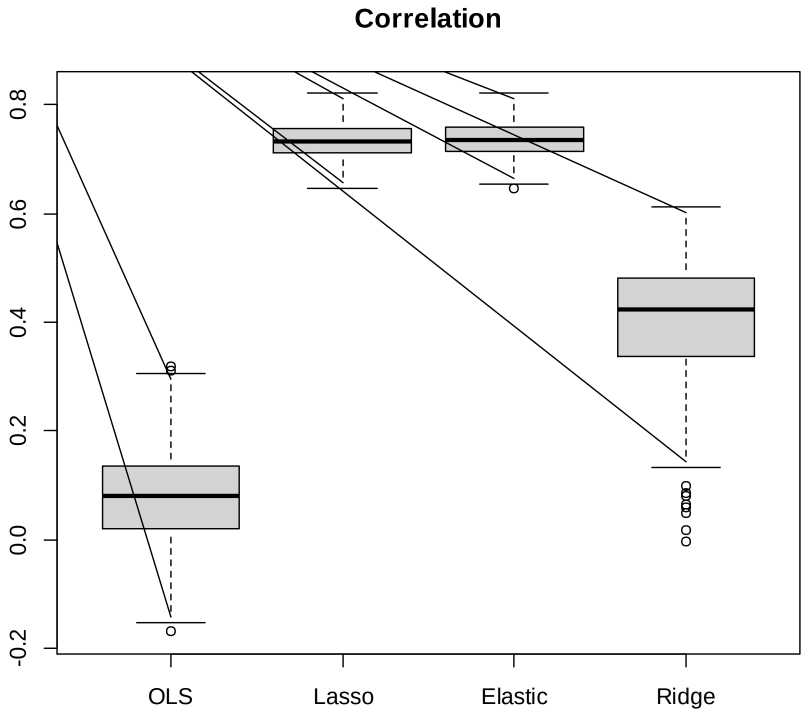
Small N, high p, Case 2

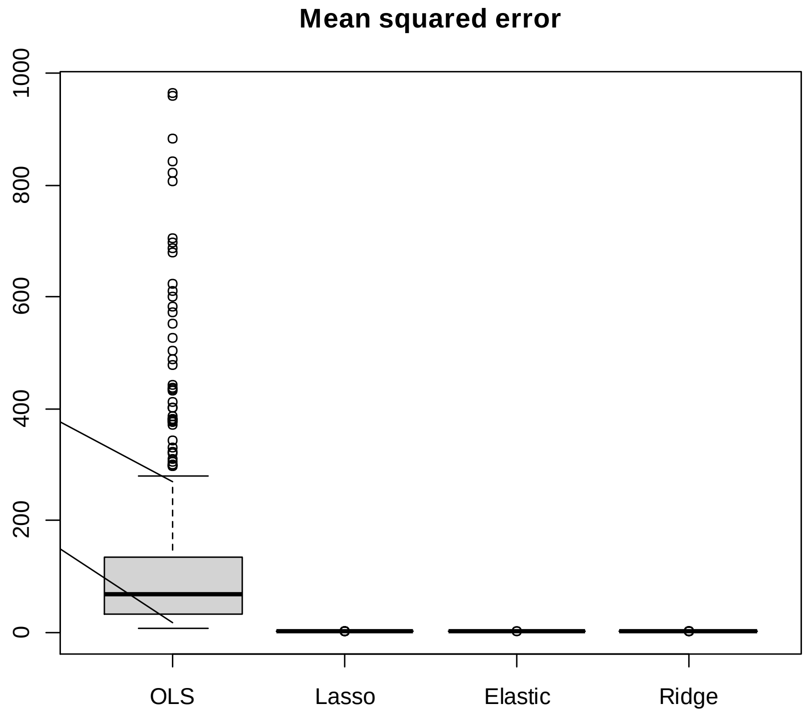
N=250, k=249, corr.y = c(0.5, 0.4, 0.3), inter-correlation r = 0.20



Small N, high p, Case 2

```
> summary(pattern.regperformance(simul.250.249, c(1, 1, 1, rep(0, times=245)) ))
    OLSp          OLSback          Lasso          Elastic          Ridge
 Mode :logical  Mode :logical  Mode :logical  Mode :logical  Mode :logical
 FALSE:500      FALSE:500      FALSE:500      FALSE:500      FALSE:500
```





Examples from more systematic simulation research

Scherr & Zhou, 2020:

- Scenarios with N=40 and up to k=100 (with two relevant predictors)
- "in cases where n > p, the OLS estimator mostly has the lowest median prediction MSE among the three different estimator scenarios – except for cases in which the number of variables is close to the number of observations"

Wester et al:

- Comparison of variable selection models for treatment X variable interactions
- combining 'lasso' and 'glinternet' led to the most accurate out-of-sample predictions, identified the most true treatment-covariate interactions, and estimated point predictions most precisely;
- but other techniques showed for example lower false positive inclusions

Hastie et al. 2017:

Compared best subset selection (based on squared errors), forward selection (based on squared errors), lasso & relaxed lasso

- best subset and lasso comparable, differential performance based on signal to noise ratio;
- best subset and forward selection very similar overall;
- the "relaxed lasso" best performing in this case.

Hastie et al., 2017, arXiv:1707.08692v2.
Scherr & Zhou, 2020, Communication Methods and Measures, 14:3, 204-211
Wester et al., 2022, *Clinical Psychological Science*. https://doi.org/10.1177/21677026211071043

Using re-sampling to evaluate stability of selection

The methods so far help identifying a subset of predictors given the selection criteria and optimisation / penalty that one selects.

But what about the stability of this selection?

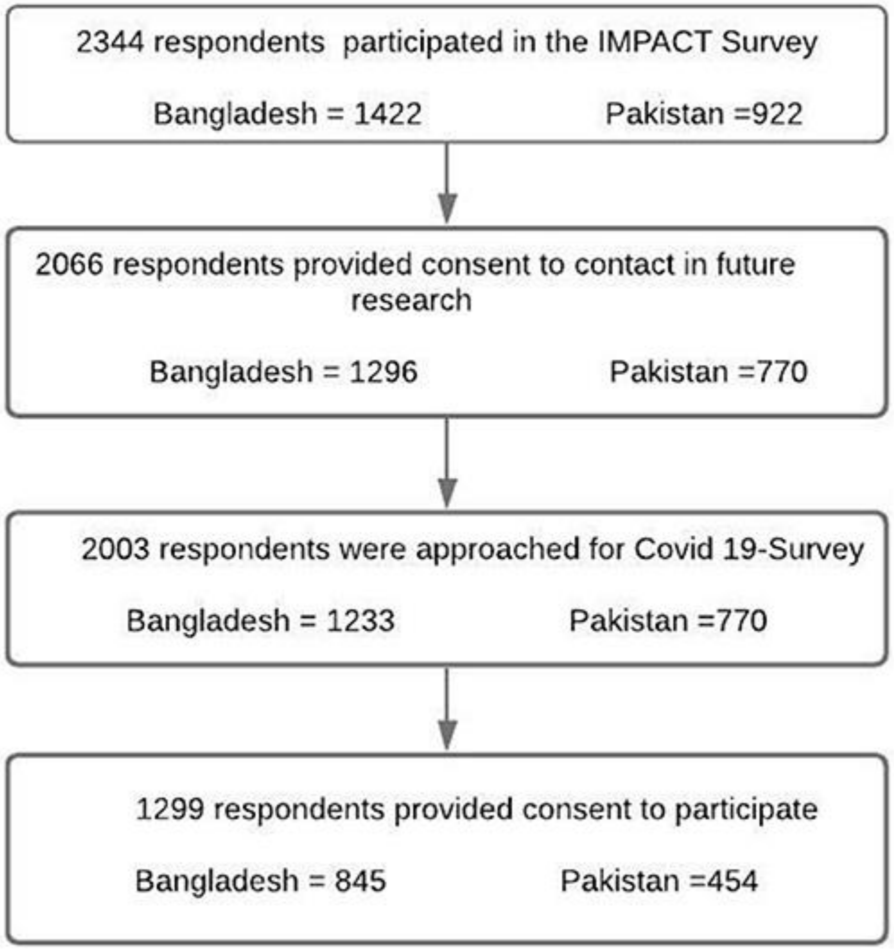
Such techniques can also be employed more formally for “model averaging”, i.e. techniques where the results of multiple/ many models are combined for better inference.

	Average Coefficient	Times selected
var2	0.32	500
var3	0.17	500
var4	0.04	349
var5	0.00	12
var6	0.00	11
var7	0.00	18
var8	0.00	70
var9	0.00	12
var10	0.00	52
var11	0.00	73
var12	0.00	19
var13	0.02	236
var14	0.01	163
var15	0.00	63

Schomaker, M., & Heumann, C. (2014). *Computational Statistics & Data Analysis, 71*, 758-770.
Wright, London & Field (2011). *Journal of Experimental Psychopathology, 2, 252-270.*
Efron, B. (1979). Annals of Statistics, 7, 1–26.

Example: Predicting COVID-related knowledge and practices

Rajan et al., 2022, *Frontiers in Psychiatry*, *13*, [785059].



Variable	"Poor" knowledge—Bangladesh		"Poor" knowledge—Pakistan		"Poor" practice—Bangladesh		"Poor" practice—Pakistan	
	OR (95% CI)	Freq.	OR (95% CI)	Freq.	OR (95% CI)	Freq.	OR (95% CI)	Freq.
Inpatient	0.97 (0.78–1.00)	215	1.09 (1.00–1.81)	239	1.01 (0.98–1.10)	113	1.00 (0.93–1.00)	50
Interview date	1.00 (1.00–1.01)	169	1.01 (1.00–1.04)	544	1.00 (1.00–1.02)	377	1.01 (1.00–1.04)	684
Monthly income (USD)	1.00 (1.00–1.01)	50	0.99 (0.90–1.00)	113	0.99 (0.85–1.00)	219	0.97 (0.80–1.00)	341
Age	1.00 (1.00–1.00)	34	1.00 (1.00–1.00)	57	1.00 (0.99–1.00)	417	1.00 (1.00–1.00)	37
Unemployed	1.05 (1.00–1.39)	305	1.00 (0.95–1.07)	79	1.00 (0.95–1.08)	101	1.00 (1.00–1.08)	56
Homemaker	1.05 (1.00–1.43)	243	1.00 (1.00–1.00)	19	0.99 (0.88–1.00)	75	0.94 (0.57–1.00)	272
Student	0.97 (0.75–1.00)	204	0.99 (0.74–1.00)	65	1.05 (1.00–1.44)	220	0.90 (0.24–1.00)	233
MINI diagnosis: major depressive disorder	1.01 (1.00–1.08)	54	0.97 (0.74–1.00)	208	0.86 (0.45–1.00)	525	1.00 (1.00–1.00)	32
MINI diagnosis: bipolar disorder with psychotic feature	1.01 (1.00–1.10)	76	0.99 (0.86–1.00)	61	0.95 (0.75–1.00)	392	1.00 (1.00–1.00)	38
Primary education	1.07 (1.00–1.42)	398	1.07 (1.00–1.55)	257	0.99 (0.90–1.00)	116	1.00 (1.00–1.01)	36
No formal education	1.05 (1.00–1.55)	191	1.02 (1.00–1.31)	121	1.01 (1.00–1.22)	115	1.02 (1.00–1.27)	91
Never married	0.98 (0.82–1.00)	162	1.06 (1.00–1.43)	265	1.00 (1.00–1.02)	58	1.06 (1.00–1.53)	240
Divorced/separated/widowed	1.07 (1.00–1.56)	270	1.08 (1.00–1.71)	220	1.00 (0.89–1.06)	89	1.02 (1.00–1.32)	79
Living urban	1.00 (1.00–1.00)	34	0.99 (0.90–1.00)	77	0.94 (0.69–1.00)	471	0.98 (0.78–1.00)	166
Home internet access	0.99 (0.91–1.00)	74	1.17 (1.00–1.73)	589	0.94 (0.69–1.00)	476	1.00 (1.00–1.00)	30
Female	1.44 (1.00–2.04)	963	0.75 (0.41–1.00)	829	0.49 (0.30–0.76)	999	0.84 (0.46–1.00)	603
SWEMWBS score	0.99 (0.97–1.00)	433	0.99 (0.95–1.00)	629	1.00 (0.98–1.00)	334	1.00 (1.00–1.00)	19
PHQ-9 score	1.00 (1.00–1.00)	14	1.00 (1.00–1.03)	306	1.00 (1.00–1.00)	30	1.00 (1.00–1.00)	19
GAD-7 score	0.99 (0.96–1.00)	303	1.00 (1.00–1.00)	22	0.96 (0.92–1.00)	954	1.00 (0.98–1.00)	117
Accessing information from social media	0.97 (0.77–1.00)	224	0.91 (0.57–1.00)	407	0.96 (0.92–1.00)	985	0.95 (0.64–1.00)	276
Accessing information from the internet	1.01 (1.00–1.14)	56	1.05 (1.00–1.65)	168	0.54 (0.31–0.85)	996	1.03 (1.00–1.39)	104
Accessing information from radio	−1.00 (0.85–1.00)	67	0.91 (0.46–1.00)	340	0.86 (0.40–1.00)	485	1.00 (1.00–1.00)	34
Accessing information from television	0.41 (0.25–0.63)	1,000	0.51 (0.27–1.00)	952	0.67 (0.43–1.00)	968	0.79 (0.44–1.00)	750
Accessing information from a newspaper	1.00 (1.00–1.00)	38	0.7 (0.66–1.00)	193	0.92 (0.61–1.00)	492	1.01 (1.00–1.11)	48
Mobility issues	0.98 (0.78–1.00)	151	1.09 (1.00–1.49)	411	1.00 (1.00–1.05)	54	1.01 (1.00–1.14)	67
Self-care issues	1.01 (1.00–1.04)	33	1.05 (1.00–1.45)	219	1.01 (1.00–1.15)	75	1.01 (1.00–1.14)	67
Difficulty doing usual activities	0.98 (0.81–1.00)	142	0.99 (0.89–1.00)	45	0.96 (0.70–1.00)	292	1.00 (0.98–1.00)	29
Pain/Discomfort	1.00 (1.00–1.02)	35	1.03 (1.00–1.31)	197	1.35 (1.00–1.97)	886	1.25 (1.00–2.03)	691
EQ-5D-VAS score	0.97 (0.96–0.98)	1,000	1.00 (1.00–1.00)	57	0.99 (0.98–1.00)	938	1.00 (1.00–1.00)	61
Poor knowledge of COVID-19 prevention measures					1.16 (1.00–1.66)	665	5.22 (2.72–8.65)	1,000

OR, Odds ratios (95% confidence intervals) from 1,000 bootstrap model runs (confidence interval determined across all bootstrap estimations incl. those where the coefficient was shrunk to 0); Freq indicating frequency of selection of this coefficient as non-zero (37); inpatient status, reference category "outpatient;" employment status, reference category "currently employed;" MINI diagnosis, reference category "non-affective psychosis;" education level, reference category "secondary/higher education;" marital status, reference category "currently married;" female, reference category male; urban living, reference category "rural".

Example: Predicting COVID-related knowledge and practices

Variable	"Poor" knowledge—Bangladesh		"Poor" knowledge—Pakistan		"Poor" practice—Bangladesh		"Poor" practice—Pakistan	
	OR (95% CI)	Freq.	OR (95% CI)	Freq.	OR (95% CI)	Freq.	OR (95% CI)	Freq.
Inpatient	0.97 (0.78–1.00)	215	1.09 (1.00–1.81)	239	1.01 (0.98–1.10)	113	1.00 (0.93–1.00)	50
Interview date	1.00 (1.00–1.01)	169	1.01 (1.00–1.04)	544	1.00 (1.00–1.02)	377	1.01 (1.00–1.04)	684
Monthly income (USD)	1.00 (1.00–1.01)	50	0.99 (0.90–1.00)	113	0.99 (0.85–1.00)	219	0.97 (0.80–1.00)	341
Age	1.00 (1.00–1.00)	34	1.00 (1.00–1.00)	57	1.00 (0.99–1.00)	417	1.00 (1.00–1.00)	37
Unemployed	1.05 (1.00–1.39)	305	1.00 (0.95–1.07)	79	1.00 (0.95–1.08)	101	1.00 (1.00–1.08)	56
Homemaker	1.05 (1.00–1.43)	243	1.00 (1.00–1.00)	19	0.99 (0.88–1.00)	75	0.94 (0.57–1.00)	272
Student	0.97 (0.75–1.00)	204	0.99 (0.74–1.00)	65	1.05 (1.00–1.44)	220	0.90 (0.24–1.00)	233
MINI diagnosis: major depressive disorder	1.01 (1.00–1.08)	54	0.97 (0.74–1.00)	208	0.86 (0.45–1.00)	525	1.00 (1.00–1.00)	32
MINI diagnosis: bipolar disorder with psychotic feature	1.01 (1.00–1.10)	76	0.99 (0.86–1.00)	61	0.95 (0.75–1.00)	392	1.00 (1.00–1.00)	38
Primary education	1.07 (1.00–1.42)	398	1.07 (1.00–1.55)	257	0.99 (0.90–1.00)	116	1.00 (1.00–1.01)	36
No formal education	1.05 (1.00–1.55)	191	1.02 (1.00–1.31)	121	1.01 (1.00–1.22)	115	1.02 (1.00–1.27)	91
Never married	0.98 (0.82–1.00)	162	1.06 (1.00–1.43)	265	1.00 (1.00–1.02)	58	1.06 (1.00–1.53)	240
Divorced/separated/widowed	1.07 (1.00–1.56)	270	1.08 (1.00–1.71)	220	1.00 (0.89–1.06)	89	1.02 (1.00–1.32)	79
Living urban	1.00 (1.00–1.00)	34	0.99 (0.90–1.00)	77	0.94 (0.69–1.00)	471	0.98 (0.78–1.00)	166
Home internet access	0.99 (0.91–1.00)	74	1.17 (1.00–1.73)	589	0.94 (0.69–1.00)	476	1.00 (1.00–1.00)	30
Female	1.44 (1.00–2.04)	963	0.75 (0.41–1.00)	829	0.49 (0.30–0.76)	999	0.84 (0.46–1.00)	603

Rajan et al., 2022, *Frontiers in Psychiatry*, *13*, [785059].

Extensions and Discussion

Regularized Structural Equation Modelling

Structural equation modelling only a very general form or regression models…

…with different deviance functions.

Jacobucci et al. (2016). *Struct Equ Modeling, 23, 555-566.*
Li, Jacobucci, Ammerman, 2021, Tutorial on the Use of the regsem Package in R. Psych. 2021; 3(4):579-592.

Variable Selection in Structural Equation Models with Regularized MIMIC Models

MIMIC models are commonly used to estimate the joint influence of a set of (presumed causal) influences on one or more latent variables.

RegSEM allows to combine confirmatory aspects of SEM with an exploratory search for important predictors.

The confirmatory and exploratory aspects can take place in either the measurement or the structural parts of a structural equation model.

Jacobucci, Brandmaier & Kievit, 2019, *Adv Methods Pract Psychol Sci, 2*, 55-76.
Jacobucci et al. (2016). *Struct Equ Modeling, 23, 555-566.*

Where epistemic goals of prediction and explanation clash

Using regularization methods to identify a set of plausible mediators is trying to combine two very different worlds:

- Mediation analysis is a causal model and relies on knowledge and interpretation of aspects such as sequence, hierarchy, and interrelationships;
- and causal variables do not necessarily have strong predictive power.

Serang et al., 2017, *Structural Equation Modeling, 24*, 733-744.
van Kesteren & Oberski, 2019, Structural Equation Modeling, 26, 710-723.

Photo by Gareth Harrison on Unsplash

Regularisation in Item Response Models

Similarly to several types of analyses in SEM / CFA also Item Response Models have a number of exploratory questions for which data need to be interrogated routinely.

A key point is the fairness of the items in a test, which is often assessed by investigating Differential Item Functioning (DIF):

Whether two (or more) groups of respondents can be characterised as showing the same response probabilities when controlled for their levels in the respective trait.

Magis et al. (2015):

Use of logistic regression models for DIF investigation and implementing this with lasso.

Tutz & Schauberger (2015) specifically for Rasch Models.

Belzak & Bauer (2020):

Similar approach, motivated by the goal to identify items that would allow anchoring / support for test equating.

Magis et al. (2015). Journal of Educational and Behavioral Statistics, 40, 111–135.
Tutz & Schauberger (2015). *Psychometrika, 80, 21-43.*
Belzak & Bauer (2020). *Psychol Methods, 25, 673-690.*

Why regularised regression or similar models?

There is a spectrum of what is now called “machine learning” approaches.

This one is fairly far to how we understand statistics and is essentially regression.

It is not a black box approach, i.e. models are fully understandable and explainable.

Especially in situations with many variables and few observations it may be useful to consider such an approach.

Epistemology: It has a clear criterion that operationalises why these selected predictors are interesting (cross-validation, error).

It puts the pressure on researchers to think about the validity of the model outside the sample.

When to use such an approach/ when to use a classic regression model

The classic regression model is a confirmatory approach that assesses whether the total of predictors derived from theory and tested together predict the dependent variable. The individual p-values from a regression model cannot be used for predictor selection (not corrected for multiple testing).

This means that approaches such as regularised regressions (of which the lasso is one approach) are more appropriate:

- the more exploratory your research question is (nothing known about the predictor space)
- the more important it is to derive a potential set of predictors from a larger number of such variables

And, with view to the areas in which these approaches were originally developed: the smaller your N/variables ratio is, the more likely it is that this approach is better suited than a standard regression model.

So what can these methods do for process-outcome research?

"Techniques for using multiple regression (MR) as a general variance-accounting procedure of great flexibility, power, and fidelity to reach aims in both manipulative and observational psychological research are presented." (Cohen, 1968, Psychological Bulletin, 70, 426-443/ p. 426)

They are likely appropriate in any setting where predictive power is the key epistemic goal.

- This requires a lot of attention to the way our samples are drawn as prediction makes only sense if it actually transcends the particular sample at hand.
- And the real advantages show only once the number of variables gets close to the number of observations.

In most cases:

Why not run a well-planned and well-justified regression model?

Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. *Big Data & Society, 1(1), 205395171452848.*

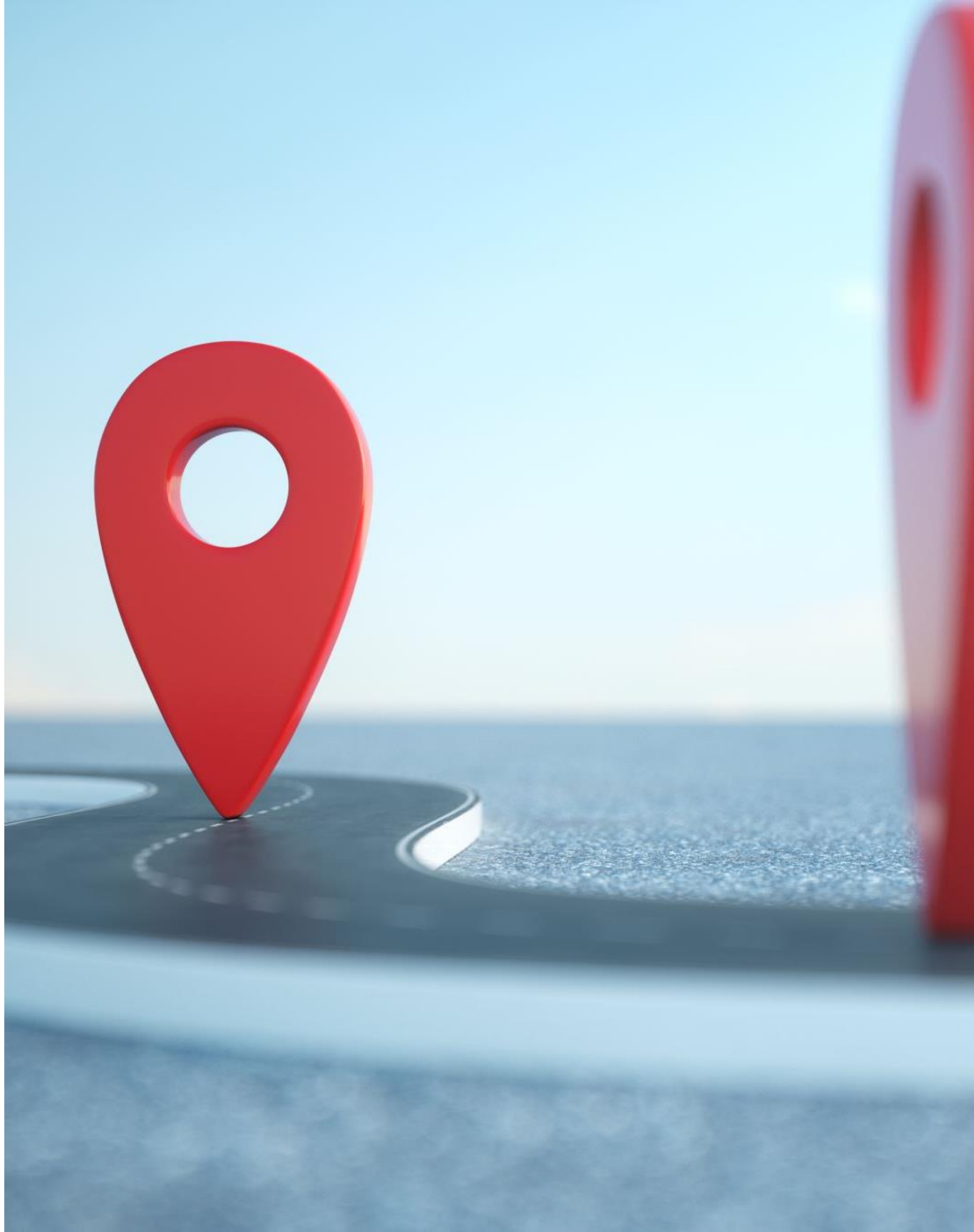
Resources

Article:

McNeish, D. M. (2015). Using lasso for predictor selection and to assuage overfitting: A method long overlooked in behavioral sciences. *Multivariate Behavioral Research*, *50*(5), 471–484.

Book:

Hastie, Tibshirani, Friedman (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.). New York: Springer.



dundee.ac.uk