Assignment 2

Martin Lindqvist

2024-02-25

Introduction

In this assignment data will be generated from a known distribution to explore and compare the efficacy of various regression and feature selection methods. The methods that will be compared are: Linear Regression utilising all available features, Linear Regression using forward selection and the Bayesian Information Criterion (BIC) for feature selection, Lasso Regression using cross-validation for feature selection, Principal Component Regression (PCR) using cross-validation for component selection, and PCR with two components.

Introduction to Simulated Dataset

A dataset has been constructed where certain features exhibit correlation, and there exists a linear relationship between a subset of the features and the response variable. The data was created in the following steps:

Parameter Definition

The fundamental parameters for the data generating process are defined as:

- $n_{\text{obs}} = 100$: The total number of observations in each simulated dataset.
- $n_x = 40$: The number of features included.
- $\rho = 0.8$: The covariance assigned to a specific subset of features.

Covariance Matrix and Feature Generation

A covariance matrix, $\mathbf{C}$, of dimensions $n_x \times n_x$, is created to control the variance and covariance among the features:

All values in the matrix are initially set to zero, implying no correlation. Features 10 to 30 ($\mathbf{C}_{10:30,10:30}$), are set to ρ, inducing a covariance of 0.8 among them.

The diagonal elements of $\mathbf{C}$ are assigned a value of 1, giving each feature a variance of 1.

Using a multivariate normal distribution with the variance/covariance structure defined by $\mathbf{C}$, a new matrix X is generated, containing n_{obs} observations of n_x normally distributed features.

Response Variable Construction

The response variable y, is generated through a summation of the first 20 features of X for each row, plus an error term ε. Where $\varepsilon \sim N(0, 1)$.

Data Compilation

The aforementioned procedure is repeated 500 times.This yields 500 distinct datasets, each comprising 100 observations with 40 features, wherein features 10 through 30 are correlated, while feature 1 through 20 has a linear relationship with the response variable, as well as stochastic noise.

Prediction MSE between the models

The predictive performance of the different models is assessed using the Prediction Mean Squared Error

(MSE), which is a measure of a model’s accuracy in forecasting on unseen data. The Prediction MSE is calculated as follows:

Prediction MSE = $\mathrm{Var}(\varepsilon) + \mathbb{E}\left[\left(f(x) - \hat{f}(x)\right)^2\right]$.

Where:

- $\mathrm{Var}(\varepsilon)$ represents the variance of the error term, which is known to be 1.
- $\mathbb{E}\left[\left(f(x) - \hat{f}(x)\right)^2\right]$ denotes the expected value of the squared differences between the true function $f(x)$ and its estimates $\hat{f}(x)$.

In the context of the data:

- $f(x_i) = \sum_{j=1}^{20} X_{ij}$ is the sum of the first 20 features for each observation i.
    - X_{ij} denotes the value of the j:th feature in the i:th row of the matrix X.
- $\hat{f}(x_i)$ are the observed predictions on the training data.

By measuring the Prediction MSE, the aim is to evaluate the effectiveness and accuracy of the various models in their ability to discern the true relationships within the data.

Below are the Prediction MSE for the evaluated models:

```
## [1] "1. OLS (all features): 1.405"
```

```
## [1] "2. LR (forward selection/BIC): 1.2952"
```

```
## [1] "3. Lasso: 1.3227"
```

```
## [1] "4. PCR (CV-folds for num components): 1.4261"
```

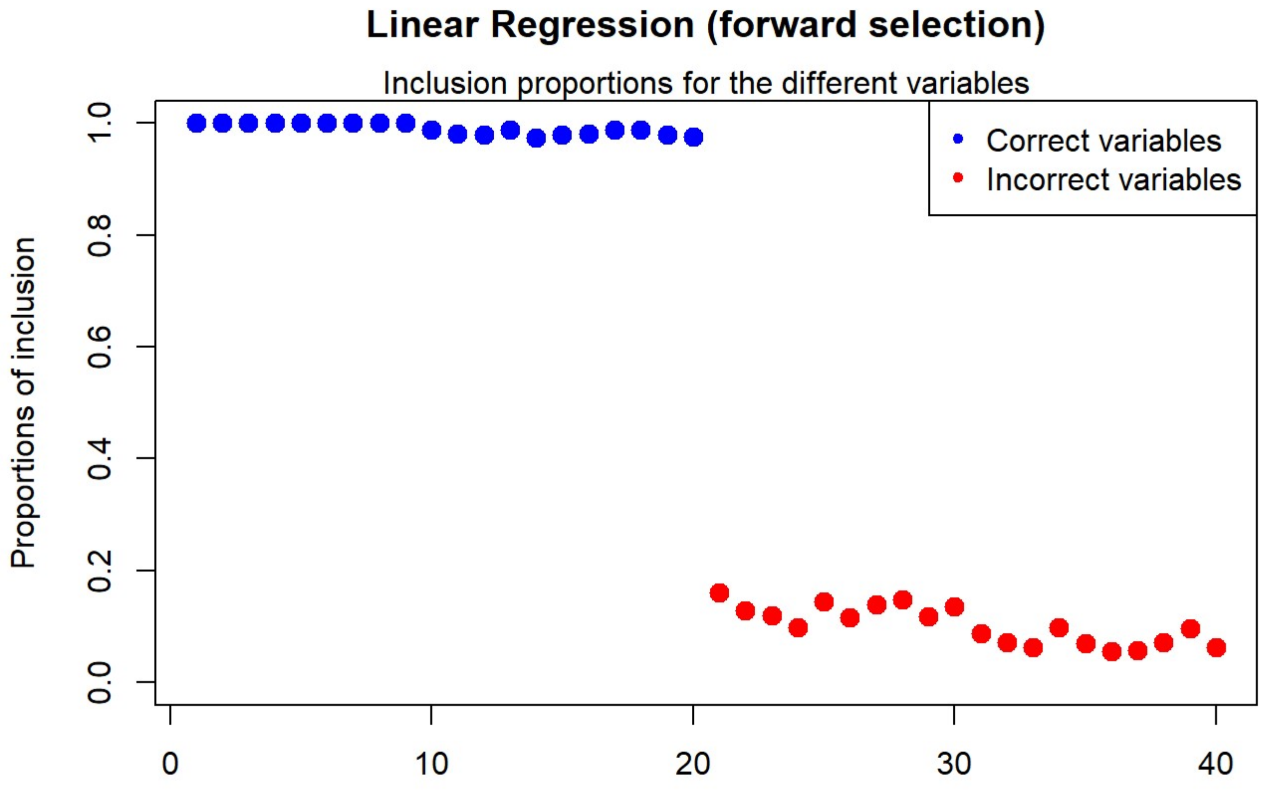
```
## [1] "5. PCR (2 components): 10.8239"
```

1. Ordinary Least Squares (OLS) with all available features

The OLS model does not perform any feature selection or dimensionality reduction. Because of this we expect the model to overfit to variables that doesn’t directly affect the response in the population (21-40). But despite this the model performs relatively well with a Prediction MSE of 1.405. This is because the linear relationship between variable 1 to 20 and the response is strong enough to be captured despite the inclusion of the other variables comprising stochastic noise.

2. Linear Regression (LR) using Forward Selection and BIC

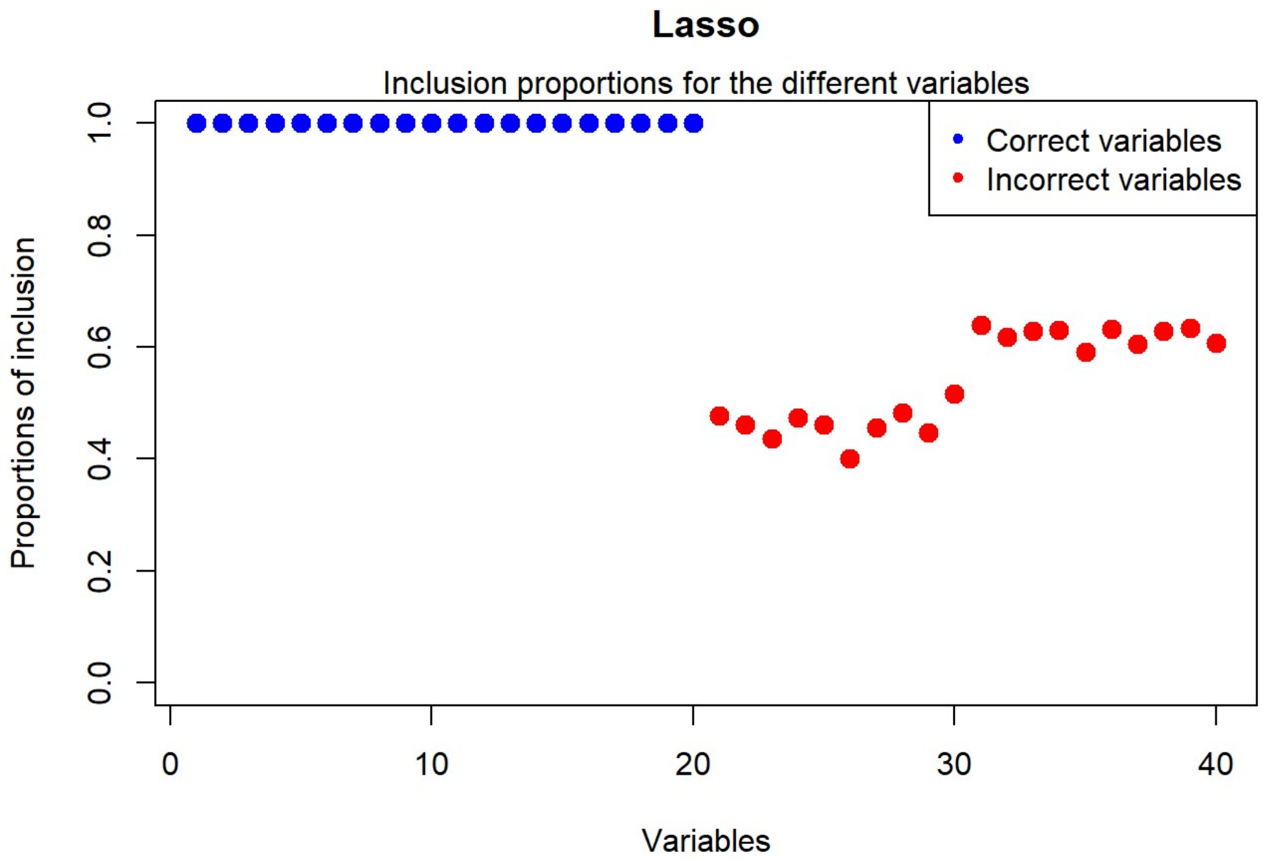
Out of all the models Linear Regression using forward selection and BIC to reduce the feature set to the most important features had the lowest Prediction MSE of 1.2952. This performance can likely be attributed to the methods ability to include the correct variables (1-20) at a very high rate, as well as including the incorrect variables (21-40) at a low rate as seen in **Plot 1**. This constructs models that often does a decent job of representing the underlying relationship within the data.



Plot 1

3. Lasso Regression

Lasso Regression using cross-validation to select the λ with the lowest cross-validation MSE has the second lowest Prediction MSE of 1.3227. It shows slightly worse performance than LR with forward selection, but better than OLS with all available features and PCR. This decrease in performance compared to LR is likely due to its inability to completely exclude the incorrect variables (21-40) as seen in **Plot 2**. We can see that features 31-40 which does not have a correlation to any other features are included at a higher rate than features 21-30 which are correlated to a subset of the correct variables.



Plot 2

4. Principal Component Regression (PCR) using Cross-Validation

PCR using Cross-Validation to select the number of components to include underperformed compared to OLS (all features), LR (forward selection) and Lasso, with a Prediction MSE of 1.4261. I had expected this model to perform better than it did. It seems like it is not able to detect the the underlying relationship in the data while simultaneously excluding the signals from the stochastic noise with the same accuracy as the other models. It does not reduce the dimensionality by much since the average number of components are 36.486 as seen below. Even though the underlying data only contains 20 variables that has an direct impact on the response.

```
## [1] "Average number of components in PCR: 36.486"
```

5. Principal Component Regression (PCR) with Two Components

PCR with only two components performs much worse than all the other models with a Prediction MSE of 10.8239. This large difference can be explained by the dimensionality being reduced by such a drastic amount that a lot of the information is lost.

Discussion

In this assignment we examined the performance of various regression and feature selection methods. A key parameter was ρ which was set to 0.8, representing the covariance between a subset of the features. Varying ρ could provide insights into the different models abilities to adapt under different data correlations. It would be especially interesting to see if Lasso Regression would be able to exclude more of the incorrect variables if ρ was set to a lower value. The performance of PCR is also likely to differ when ρ is changed. PCR often performs well in environments where the variables has high covariance since it creates components that are uncorrelated to each other. It would also be interesting to see how the performance of the various models would change if the subset of features that have a covariance would be changed to include more or fewer features.