Assignment 3

Emre Beray Boztepe

31 05 2024

1-) Prove that the trace of the symmetric real matrix is equal to the sum of its eigenvalues (Hint : use the spectral decomposition and the circular property of the trace).

We know that for real symmetric matrix A, there exists an Orthogonal matrix P (where the columns are made of eigenvectors of A) and a diagonal matrix Λ such that:

$$A = P\Lambda P^T$$

Since P is orthogonal, we can say that:

$PP^T = P^T P = I$ which is the Identity matrix

And Λ is diagonal with eigenvalues of A $(\lambda_1, \lambda_2, \ldots, \lambda_n)$:

$$\Lambda = diag(\lambda_1, \lambda_2, \ldots, \lambda_n)$$

As a property of trace operator, suppose that we have two matrices X and Y:

The trace of their product is commutative: $tr(XY) = tr(YX)$

So, when we want to apply trace operator to our spectral decomposition $A = P\Lambda P^T$:

$$tr(A) = tr(P\Lambda P^T) = tr(PP^T\Lambda) = tr(\Lambda I) = tr(\Lambda)$$

Trace of diagonal matrix is simply the sum of its diagonal elements:

$tr(\Lambda) = \lambda_1 + \lambda_2 + \ldots + \lambda_n$ which are exactly the eigenvalues of A

2-) Consider the real matrix X of the dimension n × p.

a) Prove that X'X is semipositive definite and that its eigenvalues are larger or equal to zero.

b) Prove that when p>n than at least one eigenvalue of X'X is equal to zero (i.e. X'X is singular).

Option a-)

We can say a matrix is semipositive definite if for all vectors v: $v^T A v \geq 0$:

$v^T(X^T X)v = (Xv)^T(Xv)$ for any vector $v \in \mathbb{R}^p$

We know that dot product of the vector Xv with itself is always non-negative.

$$(Xv)^T(Xv) = |\square|Xv|\square|^2 \geq 0$$

$|⬚|Xv|⬚|^2$ is sum of the squares of the components of Xv. So, $X^T X$ is semipositive definite.

Since $X^T X$ is semipositive definite, for any eigenvector v corresponding to λ:

$$v^T(X^T X)v = \lambda v^T v = \lambda|⬚|v|⬚|^2$$

And $|⬚|v|⬚|^2 > 0$ for any non-zero vector v, we can say $\lambda \geq 0$

Option b-)

We have $X_{n\times p}$ and we can say that matrix $X^T X_{p\times p}$.

When $p > n$ the rank of matrix we have $X_{n\times p}$ is at most n because n is smaller, and it is limited by the smaller dimension.

And $X^T X_{p\times p}$ has $p - n$ eigenvalues that are zero because the rank plus the dimension of the kernel of a matrix equals its total number of columns.

So, $X^T X$ is singular. Because it does not have full rank and has zero eigenvalues

3-) Your data contains 10 variables. You fit 10 regression models including the first variable, the first two variables, etc. The residual sums of squares for these 10 consecutive models are equal to (1731, 730, 49, 38.9, 32, 29, 28.5 27.8, 27.6, 26.6). The sample size is equal to 100. Which of these 10 models will be selected by AIC ? And which model will be selected by BIC or RIC? Assume that the standard deviation of the error term is known; σ = 1.

Formulas:

$$AIC = n\ln\left(\frac{RSS}{n}\right) + 2k$$

$$BIC = n\ln\left(\frac{RSS}{n}\right) + k\ln(n)$$

$$n = 100$$

$$\sigma = 1$$

```
##  [1] "AIC values: 287.128436918812"  "AIC values: 202.787434815435"
##  [3] "AIC values: -65.3349887877465" "AIC values: -86.4175935363691"
##  [5] "AIC values: -103.943428318836" "AIC values: -111.787435600162"
##  [7] "AIC values: -111.526609871349" "AIC values: -112.01341652915"
##  [9] "AIC values: -110.735441326499" "AIC values: -112.425897020044"

##  [1] "BIC values: 289.7336071048"    "BIC values: 207.997775187411"
##  [3] "BIC values: -57.5194782297822" "BIC values: -75.9969127924167"
##  [5] "BIC values: -90.917577388896"  "BIC values: -96.1564144842332"
##  [7] "BIC values: -93.290418569432"  "BIC values: -91.1720550412453"
##  [9] "BIC values: -87.2889096526059" "BIC values: -86.3741951601629"

## [1] "Model selected by AIC: 10"
## [1] "Model selected by BIC: 6"
```

4-) Assuming the orthogonal design ($X'X = I$) and n = p = 10000 calculate the expected number of false discoveries for AIC, BIC and RIC, when none of the variables is really important (i.e. p0 = p).

Formulas:

$$AIC = (p - k) \times 2 \times \left(1 - \Phi(2^{0.5})\right)$$

$$BIC = (p - k) \times 2 \times \left(1 - \Phi\left(\sqrt{log(n)}\right)\right)$$

$$RIC = (p - k) \times 2 \times \left(1 - \Phi\left(\sqrt{2log(p)}\right)\right)$$

```
## Expected Number of FD - AIC:  1572.992
## Expected Number of FD - BIC:  24.06519
## Expected Number of FD - RIC:  0.1771252
```

5-) When would you use AIC ? BIC ? RIC ?

AIC:

When we are looking for goodness of fit but not for the simplicity or interpretablity of the model.

The sample size is sufficiently large compared to the number of parameters.

BIC:

When we need model that is either good at predicting and reasonably simple.

The sample size is very large. Because BIC penalizes model complexity more heavily when sample size is large.

RIC:

When we need a balance between overfitting and underfitting.

When we need a criterion that is consistent under mild conditions (less restrictive than BIC)

Summary:

AIC is less sensitive to sample size changes. BIC and RIC are more appropriate for larger datasets. (penalty terms)

When we have a model with a large number of parameters, BIC and RIC may help prevent overfitting more effectively than AIC

For predictive accuracy, AIC might be preferable. For finding the model that balances accuracy with simplicity, BIC or RIC may be more suitable.

6-) Derive the formula for the bias, variance and mse of the ridge regression estimate under the orthogonal design (i.e when X'X=I). Compare to the respective values for the least square estimator.

For the regression model $y = X\beta + \epsilon$ where $\epsilon \sim N(0, \sigma^2 I)$, the ridge regression estimator:

$$\hat{\beta}_{ridge} = (X^T X + \lambda I)^{-1} X^T y$$

In an orthogonal design where $X^T X = I$

$$\hat{\beta}_{ridge} = (I + \lambda I)^{-1} X^T y = \frac{1}{1+\lambda} X^T y$$

Since $X^T y = X^T(X\beta + \epsilon) = I\beta + X^T\epsilon$, and given the orthogonality, $X^T\epsilon = \epsilon$ we would have:

$$\hat{\beta}_{ridge} = \frac{1}{1+\lambda}(\beta + \epsilon)$$

Bias of the ridge estimator:

$$E[\hat{\beta}_{ridge}] = E\left[\frac{1}{1+\lambda}(\beta + \epsilon)\right] = \frac{1}{1+\lambda}\beta$$

$$Bias(\hat{\beta}_{ridge}) = \frac{1}{1+\lambda}\beta - \beta = \left(\frac{1}{1+\lambda} - 1\right)\beta = -\frac{\lambda}{1+\lambda}\beta$$

Variance of the ridge estimator:

$$Var(\hat{\beta}_{ridge}) = Var\left(\frac{1}{1+\lambda}(\beta + \epsilon)\right) = \left(\frac{1}{1+\lambda}\right)^2 Var(\epsilon)$$

We know that: $Var(\epsilon) = \sigma^2 I$

$$Var(\hat{\beta}_{ridge}) = \left(\frac{1}{1+\lambda}\right)^2 \sigma^2 I$$

Mean Squared Error (MSE) of the ridge estimator:

$$MSE(\hat{\beta}_{ridge}) = Bias(\hat{\beta}_{ridge}) + Var(\hat{\beta}_{ridge})$$

$$MSE(\hat{\beta}_{ridge}) = \left(\frac{1}{1+\lambda}\right)^2 |\square|\beta|\square|^2 + \left(\frac{1}{1+\lambda}\right)^2 \sigma^2 I$$

Comparison with Least Squares:

$$\hat{\beta}_{LS} = (X^T X)^{-1} X^T y = X^T y$$

Bias: $E[\hat{\beta}_{LS}] - \beta = \beta - \beta = 0$ (unbiased)

Variance: $Var(\hat{\beta}_{LS}) = \sigma^2 (X^T X)^{-1} = \sigma^2 I$

MSE: $MSE(\hat{\beta}_{LS}) = Var(\hat{\beta}_{LS}) = \sigma^2 I$ Since it is unbiased, MSE simplifies to the variance.

7-) For a given data set with 40 explanatory variables the residual sums of squares from the least squares method and the ridge regression are equal to : 4.5 and 11.6, respectively. For the ridge regression the trace of $X(X'X+\gamma I)^{-1}X'$ is equal to 32. Which of these two methods yields the better estimated prediction error.

We have the information:

- RSS for Least Squares: 4.5
- RSS for Ridge Regression: 11.6
- Trace of $X(X'X+\gamma I)^{-1}X'$: 32

When we compare these two methods by their RSS values, we can say least squares with RSS of 4.5 fits the data better than the ridge regression model which has RSS of 11.6. Because lower RSS indicates a better fit.

When we check trace, we can say 32 out of 40 explanatory variables implies that the model still retains a significant amount of the complexity of the data despite the regularization.

As a result, least squares should suggest a better model fit in terms of minimizing the prediction error on the given dataset with having lower RSS.

8-) Given X'X=I calculate the expected value of false discoveries and the power of LASSO

Variables for LASSO are chosen according to whether or not their coefficients stay non-zero after being decreased towards zero. With the assumption that each coefficient is evaluated against a threshold given by λ (tuning parameter), the FDR may be approximately computed.

In an orthogonal design, the LASSO problem for each coefficient β_i:

$$\widehat{\beta_\iota} = sign(z_i)max(0, |z_i| - \lambda)$$

z_i: ith element of $X'y$

Under the null hypothesis, $z_i \sim N(0, \sigma^2)$

If $|z_i| > \lambda$ we can say variable is false discovered.

P(False Discovery) = $P(|z_i| > \lambda)$ = $2P(z_i > \lambda)$ = $2\left(1-\phi\left(\frac{\lambda}{\sigma}\right)\right)$

ϕ: cdf of the standard normal distribution. Expected number of FD, p × P(False Discovery), where p is the number of predictors.

Power of LASSO:

For any non-zero coefficient β_i,

$$z_i = \beta_i + \epsilon_i$$

where $\epsilon \sim N(0, \sigma^2)$

P(Power) = $P(|\beta_i + \epsilon_i| > \lambda)$

$\beta_i + \epsilon_i \sim N(\beta_i, \sigma^2)$, then:

P(Power) = $1 - P(|N(\beta_i, \sigma^2)| \leq \lambda)$

9-) Consider adaptive LASSO with $\lambda_i = w_i \lambda$.

i) How can you calculate adaptive LASSO estimator using the numerical solver for LASSO (like glmnet).

ii) In the orthogonal case $(X'X = I)$ calculate the value of the adaptive LASSO estimator for the specific coordinate of the beta vector.

iii) The ordinary least squares estimator of β_1 under the orthogonal design (X'X=I) is equal to 3 and the LASSO estimator of this parameter is equal to 2. What is the value of the adaptive LASSO estimator of β_1 if we use the same value of λ and the weight for X_1 is w_1 = 1/4.

Option i)

The adaptive LASSO problem can be solved using glmnet:

$$AL = glmnet(X, Y, alpha = 1, penalty.factor = w)$$

Option ii)

The adaptive LASSO estimator for a coefficient β_i can be simplified to:

$$\hat{\beta}_i^{AL} = sgn(\hat{\beta}_i^{LS}) max(|\hat{\beta}_i^{LS}| - w_i\lambda, 0)$$

- $\hat{\beta}_i^{AL}$ = Adaptive LASSO estimator of β_i
- $\hat{\beta}_i^{LS}$ = Least Squares estimator of β_i

Option iii)

We have:

- $\hat{\beta}_i^{LS} = 3$
- $\hat{\beta}_i^{LASSO} = 2$
- $w_1 = \frac{1}{4}$

It is wanted to calculate β_1

First, from this calculation $\hat{\beta}_i^{LASSO} = 2$, it is implied that λ was set such that $max(3 - \lambda, 0) = 2$ in the simple LASSO formula. So, we can say $\lambda = 1$

From this formula,

$$\hat{\beta}_i^{AL} = sgn(\hat{\beta}_i^{LS})max(|\hat{\beta}_i^{LS}| - w_i\lambda, 0)$$

$$\hat{\beta}_1^{AL} = sgn(3)max\left(3 - \frac{1}{4} \times 1,0\right) = max(3 - 0.25,0) = 2.75$$

So, $\hat{\beta}_1^{AL} = 2.75$

Project 1: James-Stein estimator and Prediction Error in Multiple Regression

1-) The data set Lab3.Rdata contains the matrix xx with expressions of 300 genes for 210 individuals.

Option a)

```
## First values of standardized data:
##  10.94778 8.590452 8.769656 9.42916 9.583878
```

Option b)

```
## First values of centered data:
##  0.9477803 -1.409548 -1.230344 -0.5708402 -0.416122
```

Option c)

For gene expression data that has been standardized and centered, σ^2 is typically assumed to be 1 because the data has already been scaled to have a standard deviation of 1. So, we can choose it as 1.

JS-Shrinks Zero:

$$\hat{\theta}_{JS} = \max\left(0,1 - \frac{(p-3)\sigma^2}{\sum(\text{estimate})^2}\right) \times \text{estimate}$$

JS-Shrinks Mean:

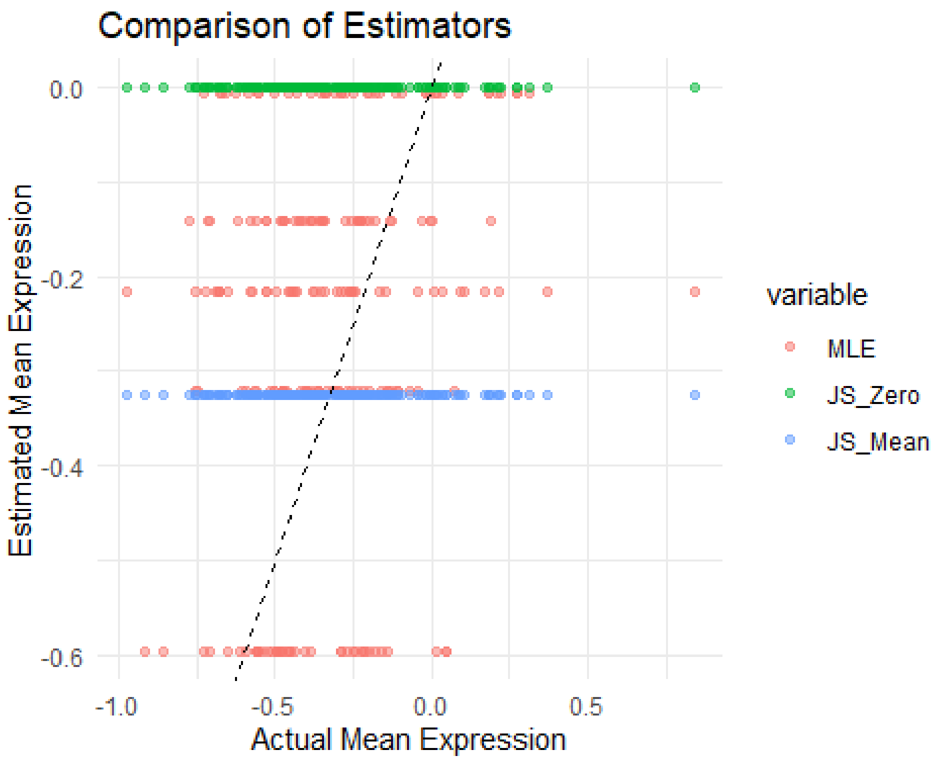
$$\hat{\theta}_{JS} = \max\left(0,1 - \frac{(p-3)\sigma^2}{\sum(\text{estimate} - \mu)^2}\right) \times (\text{estimate} - \mu) + \mu$$

- μ: overall mean
- estimate: I've used MLE means in this context
- p: length of estimate

```
## MLE Estimator:  -0.5959199 -0.005803216 -0.3205857 -0.2148937 -0.1413068
## JSE: Shrinking Towards Zero:  0 0 0 0 0
## JSE: Shrinking Towards Mean:  -0.3247122 -0.3247122 -0.3247122 -0.3247122
-0.3247122
```

Option d)

Option d Option i)



Option d Option ii)

```
## Squared Error for MLE:  18.76106
## Squared Error for JSE Zero:  36.20208
## Squared Error for JSE Mean:  14.3632
```

Option d Option iii)

Among this three, JSE Shrunk towards to mean has the lowest squared error, where MSE has a moderate squared error and JSE shrunk toward to zero has the highest one. While MLE provides a low-bias estimate, its variance might not always be optimal. JSE techniques aim to balance this issue by introducing some bias through shrinkage but not significantly reducing variance. Shrinkage towards common mean reduces the risk of extreme values driven by random sample variability.

2-) Generate the design matrix $X_{1000\times950}$ such that its elements are iid random variables from N (0, $\sigma = \sqrt{\frac{1}{1000}}$). Then generate the vector of the response variable according to the model Y = Xβ + ϵ, where $\beta = (3,3,3,3,3,0,\ldots,0)^T$ and ϵ ~ N (0, I).

i) 2 first variables
ii) 5 first variables
iii) 10 first variables

iv) 100 first variables
v) 500 first variables
vi) all 950 variables.

For each of the considered models:

Option a)

```
##          PE_name num_var      RSS       PE
## 1 Least Squares       2 51.87222 2075.392
## 2 Least Squares       5 43.94012 1981.094
## 3 Least Squares      10 57.21803 1882.063
## 4 Least Squares     100 71.18319 2097.156
## 5 Least Squares     500 29.39210 1892.549
## 6 Least Squares     950 39.24009 1958.865
```

Option b)

```
##   PE_name num_var      RSS       PE
## 1     RSS       2 64.12741 1964.127
## 2     RSS       5 47.87898 1947.879
## 3     RSS      10 54.78826 1954.788
## 4     RSS     100 42.34399 1942.344
## 5     RSS     500 53.67035 1953.670
## 6     RSS     950 42.43695 1942.437
```

Option c)

Formula provided in the class: Loo Cross-Validation = $\sum_{i=1}^{n}\left(\frac{Y_i-\widehat{Y_i}}{1-M_{ii}}\right)^2$

```
##   PE_name num_var      RSS       PE
## 1     LOO       2 50.23099 21380.09
## 2     LOO       5 53.44293 22481.97
## 3     LOO      10 63.15782 25933.34
## 4     LOO     100 44.07476 17933.17
## 5     LOO     500 54.72850 22852.97
## 6     LOO     950 51.95564 22567.94
```

Option d)

```
##               PE_name num_var      RSS        PE
## 1  Prediction Error       2 36.07334  1960.442
## 2     Least Squares       2 51.87222  2075.392
## 3               RSS       2 64.12741  1964.127
## 4               LOO       2 50.23099 21380.090
## 5  Prediction Error       5 45.61743  1929.758
## 6     Least Squares       5 43.94012  1981.094
## 7               RSS       5 47.87898  1947.879
## 8               LOO       5 53.44293 22481.966
## 9  Prediction Error      10 41.07914  1924.161
## 10    Least Squares      10 57.21803  1882.063
## 11              RSS      10 54.78826  1954.788
## 12              LOO      10 63.15782 25933.338
## 13 Prediction Error     100 46.75060  1902.874
## 14    Least Squares     100 71.18319  2097.156
## 15              RSS     100 42.34399  1942.344
## 16              LOO     100 44.07476 17933.170
## 17 Prediction Error     500 42.55710  1960.648
## 18    Least Squares     500 29.39210  1892.549
## 19              RSS     500 53.67035  1953.670
## 20              LOO     500 54.72850 22852.973
## 21 Prediction Error     950 64.80705  1882.264
## 22    Least Squares     950 39.24009  1958.865
## 23              RSS     950 42.43695  1942.437
## 24              LOO     950 51.95564 22567.941
```

Least Squares: This estimator performs best with a moderate to high number of variables, achieving its lowest PE at 10 variables (1882.063) and still performing reasonably well at 500 variables (1892.549).

RSS: Consistently produces lower PEs across all variable counts compared to Least Squares, except at 10 variables where Least Squares slightly outperforms. RSS shows a more stable PE as the number of variables increases, suggesting robustness.

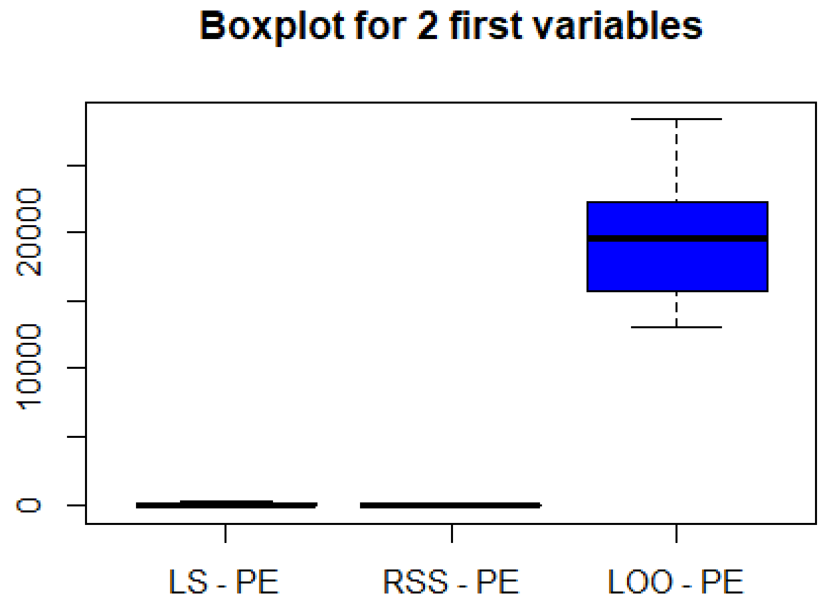
LOO: Shows significantly higher PEs across all counts of variables, suggesting it might not be as effective as the other two estimators for this particular data set.

The model using Least Squares with 10 variables appears to offer the best PE among the three estimators and is the most favorable in this analysis.

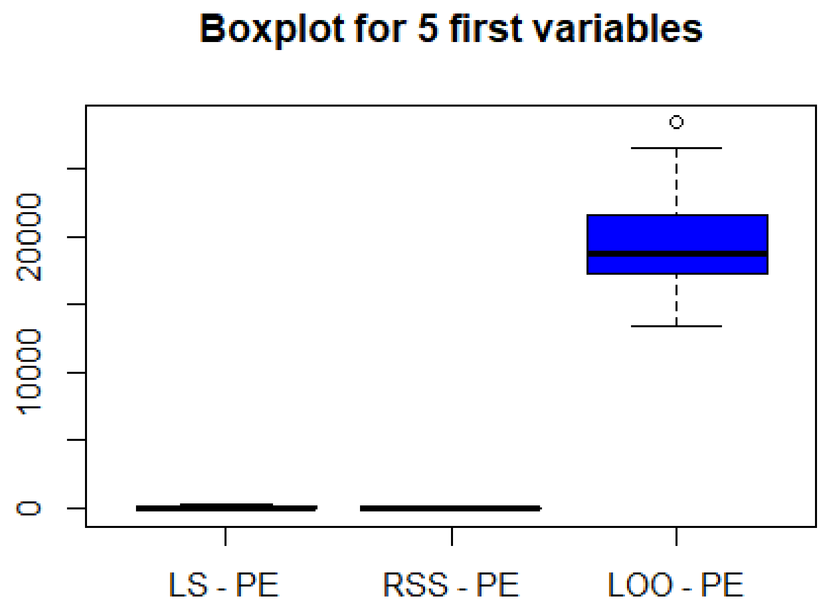
Option e)

I have a problem with plotting. As can be seen, all values for both LS - PE and RSS - PE are between -200 and 300 for all first i (i = 2, 5, 10, ..., 950) variables. But whatever I do, I could not limit y-values starting from minus values. Since all plot starts in y axis from 0, it looks like all differences are 0 (like RSS, PE and LS are the same) they are not as can be seen from the results that are printed. Also, having LOO around 20000 is a big difference when we compare with other estimators.

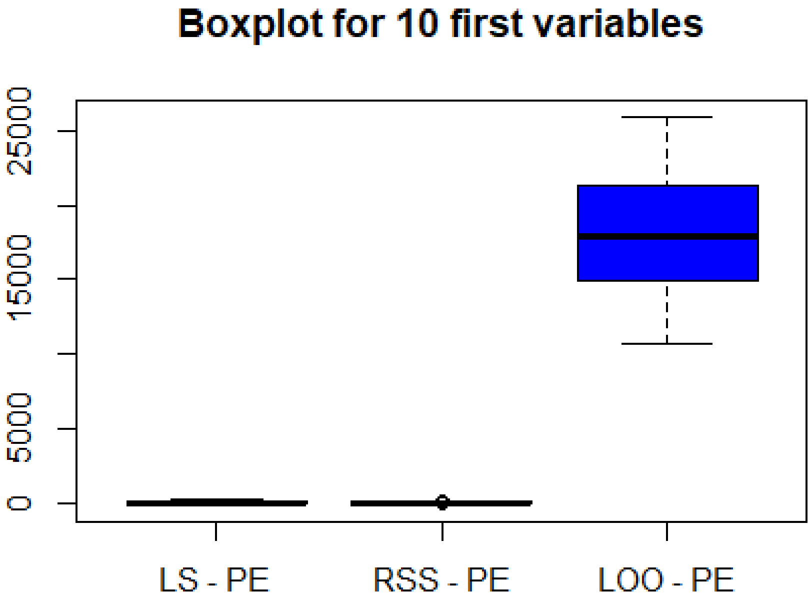
```
## [1] "Results for 2 first variables"
##    Least Square        RSS      LOO
## 1    118.128056  91.706097 24273.65
## 2     70.852603  95.383742 25178.36
## 3     14.081174  21.062694 20029.75
## 4   -171.079576 -46.348364 21902.08
## 5    -70.857739  32.191327 21296.93
## 6      1.173392 -10.095907 17733.15
## 7    -90.040880 -15.799065 24836.99
## 8    -53.056366 -45.241730 20853.85
## 9    127.845179 -48.780264 21115.70
## 10   112.026968  19.178564 16481.12
## 11  -120.326319 -93.010403 18289.35
## 12   -22.898876  14.279171 22742.06
## 13    52.334236  25.019182 13680.27
## 14     6.752555 -17.874482 22274.02
## 15  -168.171752 -64.775462 28371.53
## 16   -37.405391   3.914876 15808.27
## 17  -121.450774 -37.980650 15654.21
## 18    44.410328   7.863698 15763.46
## 19   -32.464474 -24.822023 20910.70
## 20   -88.580037  94.285599 13765.54
## 21   209.818519  31.388795 13028.96
## 22    23.188333  21.338202 23210.55
## 23    14.499311 -23.728792 15785.20
## 24  -127.068776 -69.108957 19250.11
## 25    57.490865 -15.968981 13630.95
## 26   132.517513  50.246813 19755.01
## 27    -3.039177  64.458638 14092.93
## 28   105.149805 -17.326230 19355.19
## 29   -28.757547  14.938031 23639.43
## 30    36.775176  -8.670828 13670.67
```



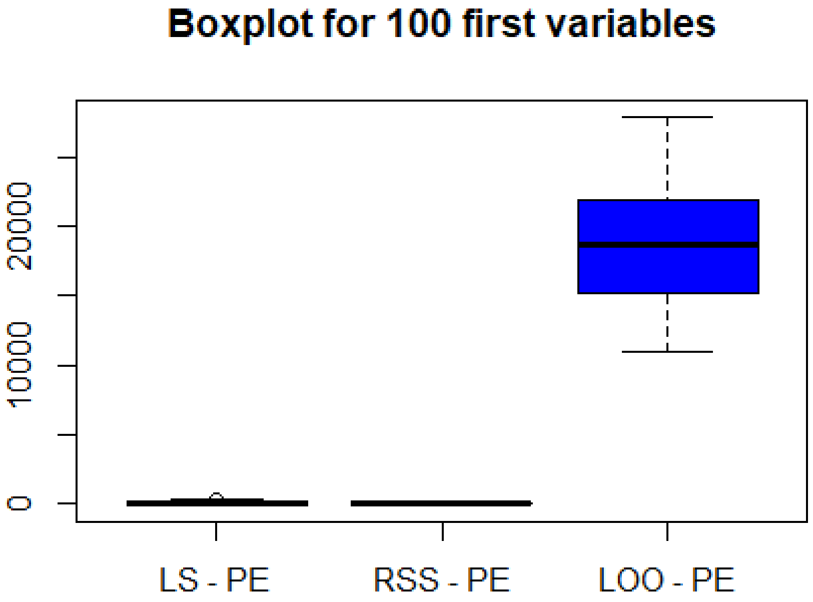
```
## [1] "Results for 5 first variables"
##    Least Square        RSS      LOO
## 1    207.190252 127.104193 13434.43
## 2   -170.475141 -33.772509 19067.35
## 3    -29.707192 -64.597832 17448.48
## 4   -117.947114 -16.908690 19399.41
## 5     18.180768 -17.700499 15768.11
## 6     75.764334  94.174635 26553.12
## 7   -101.950191 -14.955906 20164.89
## 8     27.160294  61.828895 17271.81
## 9    -60.996572 -29.705380 15394.81
## 10    -6.165194   5.842544 20471.35
## 11   128.739253  34.844059 28487.69
## 12    41.822995 -13.812292 18479.42
## 13   -58.632898 -24.329049 21584.89
## 14   135.470725 -26.671850 23690.80
## 15    19.498941  55.668353 19596.49
## 16  -117.581350 -30.606237 22221.67
## 17   -21.166826 -76.434522 18387.89
## 18   181.868425  59.086387 17088.84
## 19   -50.150636  83.373318 18706.11
## 20    42.069777 -32.453220 17751.68
## 21  -110.390863   6.127392 18551.29
## 22   -33.409651 -31.156664 16918.81
## 23    88.051371  13.387340 19459.70
## 24   -90.250780  53.731751 23385.58
## 25    17.541888  -6.071286 18135.37
## 26   -14.544253 -51.459031 16611.00
## 27  -112.268918 -62.518340 16499.55
## 28    54.989298  20.630952 18875.93
## 29    47.182835 -46.408856 23881.33
## 30   -60.478716  -1.125716 21768.95
```



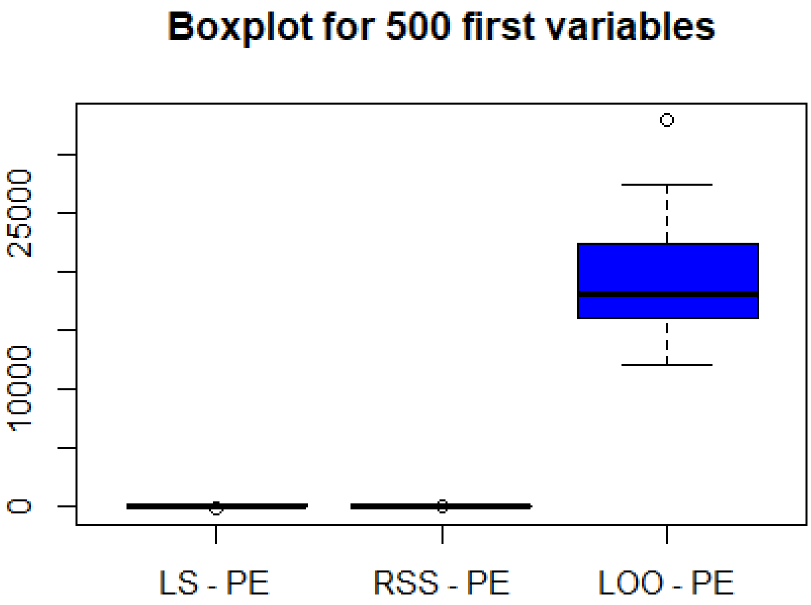
```
## [1] "Results for 10 first variables"
##    Least Square         RSS      LOO
## 1   -53.3847930  33.5669107 13642.83
## 2    -2.7751420   7.1413335 14859.79
## 3    20.2781294   7.6840670 18212.84
## 4   105.8664022 109.0412118 16200.16
## 5    89.3439809  22.8934303 19888.40
## 6   129.3016010  12.3954046 16339.58
## 7   171.2663423  29.5999589 18416.84
## 8    59.9461500  41.5135896 25331.36
## 9   -45.9201246  -4.9258609 14085.07
## 10 -140.6158623 -87.1426871 11269.95
## 11  152.2011258  20.6020160 17355.31
## 12  100.2856469  32.5651193 15331.84
## 13 -158.7904400 -22.7874204 17672.36
## 14   91.8058342  48.4337014 21339.20
## 15  -81.6543707  10.6585138 13831.66
## 16  -40.4054338 -18.3041262 24202.92
## 17  -71.6061562  23.8623279 16054.07
## 18  107.6147313  32.7759947 19470.77
## 19 -142.5127408 -23.5411619 23211.48
## 20   30.0398726  -7.5612941 10673.28
## 21   56.8904194  32.9270674 21920.73
## 22    2.6162373  32.2902599 19781.15
## 23  152.8050526  37.9387881 19128.49
## 24   -0.2552031 -14.7383412 25922.88
## 25  -40.1853869   5.4738576 14825.04
## 26   24.7465370 -38.6858910 23662.47
## 27  -34.9756061  21.2141308 12588.62
## 28 -118.9675442  29.2791831 24258.67
## 29   23.3343350  48.1961627 19844.39
## 30  138.5683773   0.9620176 17237.52
```



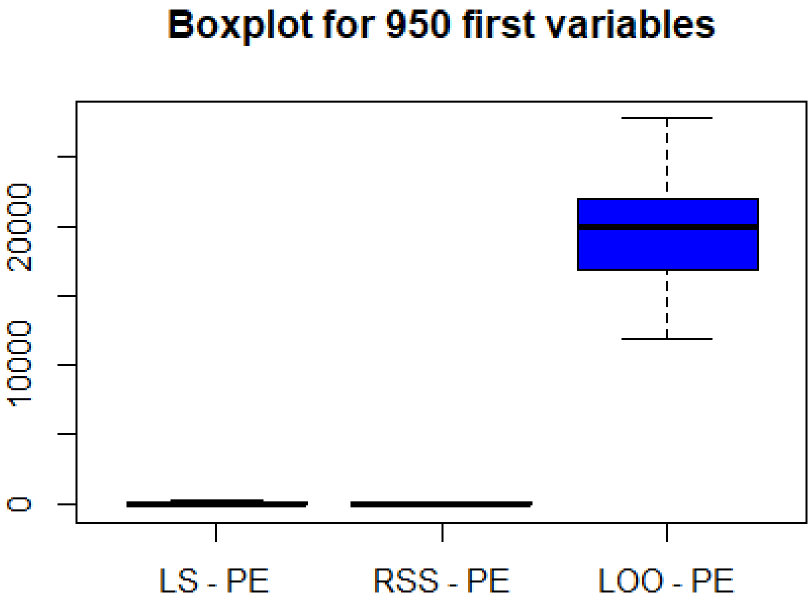
```
## [1] "Results for 100 first variables"
##    Least Square          RSS      LOO
## 1    -12.286820 -13.7951802 18630.58
## 2   -121.709450 -24.2789855 19002.97
## 3   -116.118732  14.5100219 24956.12
## 4     52.460287 -51.8986867 17537.97
## 5    129.232513 -47.1870865 22409.94
## 6    192.757556  24.5778027 20870.50
## 7    -95.738921  33.5832467 21948.78
## 8    174.233550   3.8437071 14882.51
## 9   -139.427944 -27.1314160 17951.98
## 10   125.531654 -71.6541324 13871.52
## 11    27.240567   7.3268469 14580.23
## 12   -87.190208  38.5942638 17802.59
## 13   -84.124793 -72.0806576 16149.86
## 14     7.968272  23.4919288 21261.44
## 15   -96.671470  -4.6182996 16096.66
## 16    44.161124 -10.0635590 16427.42
## 17   311.528863  60.2717537 18821.63
## 18    22.638235 -38.2819083 10981.40
## 19   -22.367931 -32.5837308 20984.50
## 20   -53.980908  26.1097950 20777.03
## 21    31.643312  50.5945359 15253.25
## 22   244.552276  39.7914130 23982.32
## 23   -56.409280  -5.3247836 14143.95
## 24  -127.834736 -72.6853605 27937.29
## 25    28.001694  18.6572451 24252.62
## 26    -1.940554  -0.9168073 13609.01
## 27    68.781546  -4.1736821 20197.38
## 28    71.738430 -21.6472200 22662.95
## 29    57.170064 -18.8062345 15012.69
## 30   -85.086211  61.1361823 23935.16
```



```
## [1] "Results for 500 first variables"
##    Least Square         RSS      LOO
## 1     16.755050 -22.762784 24575.74
## 2     21.216090 -28.746991 20724.48
## 3     11.555886  16.223204 21897.19
## 4    -17.979444  51.554757 15949.50
## 5    -24.014210  21.826599 16590.14
## 6    -36.064345  16.324029 16867.61
## 7      5.224006  25.722820 23634.40
## 8    142.414090  64.880046 27352.42
## 9   -203.959477  16.662739 26389.91
## 10    72.063895   9.687438 20772.61
## 11    82.815323  22.903577 18652.04
## 12    62.500512  28.567877 20197.38
## 13    54.072478   3.589299 13968.71
## 14    30.984104  19.823608 15160.89
## 15   129.597065  33.578829 24030.84
## 16    63.386780  21.496378 24716.58
## 17   -94.010237   4.389608 17104.57
## 18     2.473365  74.386452 32935.32
## 19    93.840178  51.446328 16404.75
## 20   -90.802469 -51.522508 12076.11
## 21   -76.868957  -9.764223 15978.76
## 22   133.932957  15.223116 21218.93
## 23    78.682694 -17.079787 17221.89
## 24   -43.730284 -88.651125 16974.63
## 25    40.843943  12.917212 19405.14
## 26   -85.043759 -47.340005 12991.59
## 27    -6.982979 -18.652746 17370.64
## 28   -39.179052  44.905850 22418.11
## 29   107.353755 -45.787759 15514.97
## 30   -28.686771  34.756014 12406.74
```



```
## [1] "Results for 950 first variables"
##    Least Square        RSS      LOO
## 1      5.300869  -4.129964 21891.47
## 2    -32.120938  -8.785205 20677.22
## 3     -2.015735  31.490502 17021.83
## 4    -68.508039 -36.085109 21782.77
## 5     98.928731  40.778012 16897.64
## 6   -182.547260   4.243991 16794.53
## 7   -161.433378 -95.836332 22008.90
## 8    -17.242254 -35.218993 27829.32
## 9    -31.609024  13.603800 17798.97
## 10   -64.239395 -21.180462 12803.95
## 11    17.744235  13.201483 12020.62
## 12   192.857448  30.617333 20309.13
## 13  -155.903663 -78.272059 24557.54
## 14    77.082921  39.416153 16830.62
## 15   -54.873024 -13.547528 23346.97
## 16   -78.510756  34.081638 16109.57
## 17  -151.795989 -60.662164 25608.58
## 18   -80.877854 -59.256021 17542.43
## 19    61.344979  49.264862 21375.20
## 20    47.679159  44.804999 14359.66
## 21    44.200301   9.139835 11902.06
## 22   -31.662741 -39.745310 21484.69
## 23   -27.950633 -46.384004 18549.75
## 24   -43.578900 -16.945605 23362.06
## 25   -63.221761  41.543308 25050.72
## 26    85.540288  54.972932 23820.68
## 27  -104.059409 -58.712034 19351.34
## 28    91.755809  47.175072 18959.54
## 29   -60.951759 -22.840393 21953.34
## 30    27.747049 -14.347233 19481.66
```



From the plots, we cannot see any outliers (small circles in plots) when we use variables 2 and 950. Using 5 variables, there are outliers for LOO estimator. For 10 variables, there are outliers for RSS estimator. Using 100 variables, there are outliers for LS. Using 500 variables, there are outliers for each estimator.

As the number of variables increases from 2 to 950, the median and variability of the prediction errors in the LOO model generally increase. This could suggest overfitting as more variables are included, which is common in models that are too complex relative to the amount of data. Sensitivity of the LOO method to the changes in data and possibly its greater reliability in reflecting the true prediction error in these settings.

Methods RSS and LS are either not very sensitive to the changes in the data or are consistently close to perfect under the tested conditions because having values so close to real PE value that we have.

Project 2: Multiple regression - model selection and regularization

Generate the design matrix X1000×950 such that its elements are iid random variables from N $(0, \sigma = \sqrt{\frac{1}{1000}})$. Then generate the vector of the response variable according to the model

Y = Xβ + ϵ , where $\beta_1 =\ldots= \beta_k = 6, \beta_{k+1} =\ldots= \beta_p = 0$ with k = 20, and ϵ ∼ N(0, I).

Analyse this data using

- mBIC2
- Ridge
- LASSO (min, 1se)
- Tuned LASSO
- SLOPE

```
## Indices:  9 6 2 4 15 10 8 5 19 1 12 3 13 14 16 11 18 20 17 7
```

For each of these methods calculate the square estimation errors $|⬚|\beta\hat{} - \beta|⬚|^2$ and $|⬚|X(\beta\hat{} - \beta)|⬚|^2$. In case of LASSO and SLOPE consider also estimators obtained by performing the regular least squares fit within the selected model. For all methods apart from ridge calculate also the False Discovery Proportion and the True Positive Proportion (Power)

```
##              Square Estimation Error Prediction Error TPR       FDR
## mBIC2                       21.42479         21.29828   1         0
## Ridge                       541.0682         417.5794   1 0.9789474
## Lasso Min                   88.54395          83.3924   1 0.6666667
## Lasso 1se                   147.9151         145.2264   1 0.2592593
## Lasso Tuning                1510.394         613.2213   1 0.9738903
## SLOPE                       31894.09         977.2353   1 0.9789474
```

mBIC2 shows the best overall performance with the lowest errors and perfect scores in TPR and FDR, suggesting it's highly effective for both identifying relevant variables and avoiding irrelevant ones.

Ridge and Lasso methods vary in effectiveness, with some trading off between high TPR and high FDR. Higher FDR in these methods might suggest they are fitting some noise as signal.

SLOPE shows the poorest performance in terms of errors, which might indicate issues with its application to this particular dataset or perhaps its sensitivity to parameter settings.

Regular Least Squares for all LASSO models and SLOPE

```
## [1] "First 5 coefficients of regular least squares for each model: "

##                    X1       X2       X3        X4       X5
## Lasso Min    6.317529 5.280544 5.345658  6.495517 6.193235
## Lasso 1se    6.734969 5.882063 5.838075  6.911914 6.525610
## Lasso Tuning 6.734787 4.422472 7.482839  7.107239 5.396130
## SLOPE        4.538134 9.437915 4.971200 12.372389 7.505228
```

SLOPE, in particular, shows a tendency to produce larger coefficients possibly because it assigns a different penalty to each coefficient, depending on its rank among the absolute values of coefficients. A larger coefficient in one model versus another could indicate a greater perceived importance of that predictor variable under certain regularization constraints.