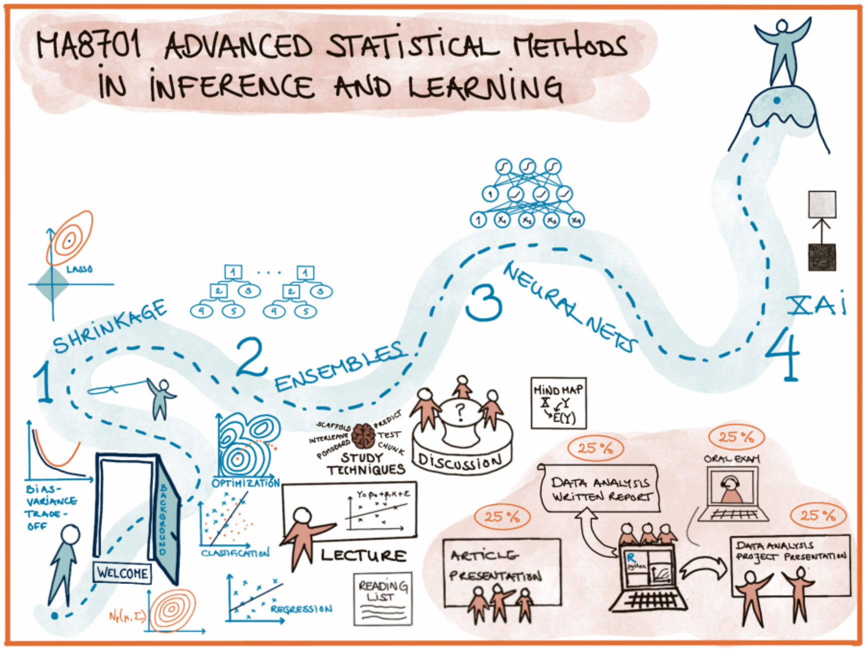
Code

MA8701 Advanced methods in statistical inference and learning

L2: Shrinkage - the beginning

Mette Langaas IMF/NTNU

15 January, 2021



Shrinkage

Literature L2

- [ELS] The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition (Springer Series in Statistics, 2009) by Trevor Hastie, Robert Tibshirani, and Jerome Friedman. Chapter 3.2 and 3.4.
- [HTW] Hastie, Tibshirani, Wainwrigh: "Statistical Learning with Sparsity: The Lasso and Generalizations". CRC press. Ebook (https://trevorhastie.github.io/). Chapter 1, 2.1-2.3,2.5.

What is in a name?

This part of the course could have been called:

- "Regularized linear and generalized linear models"
- "Penalized maximum likelihood estimation"
- and also "Sparse models",

but it is called "Shrinkage".

Focus is on generalized linear models, but we will also consider shrinkage in the next parts of this course (then for "more complex" method).

Question: in linear models (linear regression, generalized linear regression) we mainly work with methods where parameter estimates are unbiased - but might have high variance and not give very good prediction performance. Can we use penalization (shrinkage) to produce parameter estimates with some bias but less variance, so that the prediction performance is improved?

We will look at different ways of penalization (which produces shrunken estimators) - mainly what is called ridge and lasso methods.

Ridge is not a sparse method, but lasso is. In sparse statistical models a *small number of covariates* play an important role.

HTW (page 2): *Bet on sparsity principle: Use a procedure that does well in sparse problems, since no procedure does well in dense problems.*

Shrinkage (penalization, regularization) methods are especially suitable in situations where we have more covariates than observations $N << p$. Two examples are

- in medicine with genetic data, where the number of patient samples is less than the number of genetic markers studied,
- in analysis of text (more to come in L3)

Linear models

(ELS 3.2, HTW Ch 2.1)

We will only consider linear models in L2, and move to generalized linear models in L3.

Set-up

Random response Y and p-dimensional (random) covariates X.

Training data: N (independent) observations: (y_i, x_i), where x_i is a column vector with p covariates (features).

Linear regression model

(ELS 3.2)

Additive noise model

$$Y = f(X) + \varepsilon$$

with $\mathrm{E}(\varepsilon) = 0$ and $\mathrm{Var}(\varepsilon) = \sigma^2$.

With squared loss, we remember that the optimal $f(X) = \mathrm{E}(Y \mid X)$.

Linear regression model - we assumes that

$$f(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j$$

is linear in X, or that is a good approximation.

The unknown parameters are the regression coefficients $\beta_0, \ldots, \beta_p$ and the error variance σ_ε^2.

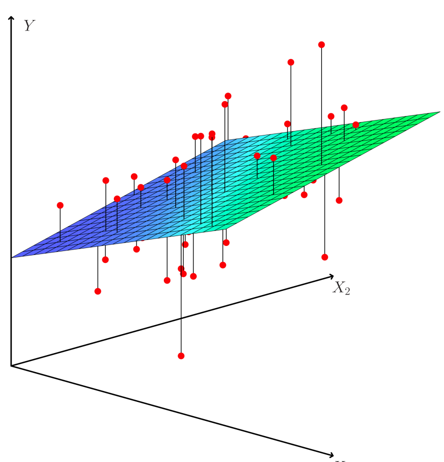


Figure from An Introduction to Statistical Learning, with applications in R (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani.

From TMA4267 we know that if (X, Y) is jointly multivariate normal, then the conditional distribution of $Y \mid X$ has mean that is linear in X and variance that is independent of X. Brush-up: See classnotes page 8 (https://www.math.ntnu.no/emner/TMA4267/2017v/TMA4267V2017Part2.pdf).

Covariates

The covariates X can be both quantitative or qualitative, be made of basis expansions or interactions - and more. For qualitative covariates often a dummy variable coding is used. Brush-up: See TMA4315 GLM Module 2 (https://www.math.ntnu.no/emner/TMA4315/2018h/2MLR.html#categorical_covariates_-_dummy_and_effect_coding).

For now we don´t say so much more, but later we want the covariates to be standardized and the reponse to be centered.

Least squares estimation

We assume that the regression parameters $\beta = (\beta_0, \beta_1, \ldots, \beta_p) \in \Re^{(p+1)}$.

We will use the word *linear predictor* $\eta(x_i) = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j$, for the linear combination in the parameters β.

The least squares estimator for the parameters β is found by minimizing the squared-error loss:

$$\text{minimize}_{\beta}\{\sum_{i=1}^{N}(y_i - \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j)^2\}$$

For derivation of the least squares estimator $\hat{\beta}$ see TMA4268 Module 3 (https://www.math.ntnu.no/emner/TMA4268/2019v/3LinReg/3LinReg.html#parameter_estimation) and links therein.

The same results are found using likelihood theory, if we assume that $Y \sim N$. See TMA4315 GLM Module 2 (https://www.math.ntnu.no/emner/TMA4315/2018h/2MLR.html#likelihood_theory_(from_b4)). Both methods are written out in these class notes from TMA4267/8 (https://www.math.ntnu.no/emner/TMA4268/2018v/notes/LeastSquaresMLR.pdf).

The squared error loss to be minimzed can be written

$$(\mathbf{Y} - \mathbf{X}\hat{\beta})^T(\mathbf{Y} - \mathbf{X}\hat{\beta})$$

The solution is

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

where $\mathbf{X}$ is a $N \times (p+1)$ matrix of covariates and $\mathbf{Y}$ is a N dimensional column vector.

Properties of regression estimators

For the classical linear model we assume

$$Y_i = \beta_0 + \sum_{j=1}^{p} X_j\beta_j + \varepsilon_i$$

with $\text{E}(\varepsilon_i) = 0$ and $\text{Var}(\varepsilon_) = \sigma_\varepsilon^2$.

This can also be written with vectors and matrices for the $i = 1, \ldots, N$ observations.

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon$$

where $\mathbf{Y}$ is a $N \times 1$ random column vector, $\mathbf{X}$ a $N \times (p+1)$ design matrix with row for observations and columns for covariates, and ε $N \times 1$ random column vector

The assumptions for the classical linear model is:

1. $\text{E}(\varepsilon) = \mathbf{0}$.
2. $\text{Cov}(\varepsilon) = \mathbf{E}(\varepsilon\varepsilon^\mathbf{T}) = \sigma^2\mathbf{I}$.
3. The design matrix has full rank, $\text{rank}(\mathbf{X}) = (p+1)$.

The classical *normal* linear regression model is obtained if additionally

4. $\varepsilon \sim N_n(\mathbf{0}, \sigma^2\mathbf{I})$ holds.

For random covariates these assumptions are to be understood conditionally on $\mathbf{X}$.

If we only assume a classical linear model, the mean and covariance of $\hat{\beta}$ is $\text{E}(\hat{\beta}) = \beta$ and $\text{Cov}(\hat{\beta}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$.

For the classical normal linear model:

- Least squares and maximum likelihood estimator for β:

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

  with $\hat{\beta} \sim N_p(\beta, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$.

- Restricted maximum likelihood estimator for σ^2:

$$\hat{\sigma}^2 = \frac{1}{n-p}(\mathbf{Y} - \mathbf{X}\hat{\beta})^T(\mathbf{Y} - \mathbf{X}\hat{\beta}) = \frac{\text{SSE}}{n-p}$$

  with $\frac{(n-p)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p}$.

- Statistic for inference about β_j, c_{jj} is diagonal element j of $(\mathbf{X}^T\mathbf{X})^{-1}$.

$$T_j = \frac{\hat{\beta}_j - \beta_j}{\sqrt{c_{jj}}\hat{\sigma}} \sim t_{n-p-1}$$

The Gauss-Markov theorem

(ELS 3.2.2)

The Gauss-Markov theorem is the famous result stating: *the least squares estimators for the regression parameters β have the smallest variance among all linear unbiased estimators*.

For simplicity, we look at a linear combination of the parameters, $\theta = a^T\beta$, with estimator $\hat{\theta} = a^T\hat{\beta} = a^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$. Observe that the estimator is linear in the response $\mathbf{Y}$.

Q: why is a linear combination of interest? What about a prediction of the response at covariate x_0? It would be $f(x_0) = x_0^T\beta$, a linear combination of the β elements.

If we assume that the linear model is correct, then $\hat{\theta}$ is an unbiased estimator of θ, because $\text{E}(a^T\hat{\beta}) = a^T\text{E}(\hat{\beta}) = a^T\beta = \theta$.

According to the Gauss-Markov theorem: if we have another estimator $\tilde{\theta} = c^T\mathbf{Y}$ that is unbiased for θ then it must have a larger variance than the LS-estimator:

$$\text{Var}(\hat{\theta}) = \text{Var}(a^T\hat{\beta}) \leq \text{Var}(c^T\mathbf{Y}) = \text{Var}(\tilde{\theta})$$

In Exercise ELS 3.3a we prove the Gauss-Markov theorem based on this set-up (least squares estimator of a linear combination $a^T\beta$).

Proof for the full parameter vector β (not only the scalar linear combination), requires a bit more work (it is ELS exercise 3.3b if you want to try).

It is not hard to check that an estimator (for example $p \times 1$ column vector) is unbiased (in each element).

Comparing variances of estimators

But, what does it mean to compare the variance (covariance matrix) of two estimators of dimension $p \times 1$?

In statistics a common strategy is to consider all possible linear combinations of the elements of the parameter vector, and check that the variance of estimator $\hat{\beta}$ is smaller (or equal to) the variance of another estimator $\tilde{\beta}$.

This is achieved by looking at the difference between the covariance matrices $\text{Cov}(\tilde{\beta}) - \text{Cov}(\hat{\beta})$. If the difference is a semi positive definite matrix, then every linear combination of $\hat{\beta}$ will have a variance that is smaller or equal to the variance of the corresponding linear combination for $\tilde{\beta}$.

Why is this correct?

Assume we want to see if $\text{Var}(c^T\tilde{\beta}) \geq \text{Var}(c^T\hat{\beta})$ for any (nonzero) vector c.

We know that $\text{Var}(c^T\hat{\beta}) = c^T\text{Cov}(\hat{\beta})c$ and $\text{Var}(c^T\tilde{\beta}) = c^T\text{Cov}(\tilde{\beta})c$.

We then consider

$$\text{Var}(c^T\tilde{\beta}) - \text{Var}(c^T\hat{\beta}) = c^T(\text{Cov}(\tilde{\beta}) - \text{Cov}(\hat{\beta}))c$$

If $\text{Cov}(\tilde{\beta}) - \text{Cov}(\hat{\beta})$ is semi positive definite then the variance difference will be equal or greater than 0 - by the definition of a semi positive definite matrix.

Mean squared error

We want to study the mean squared error for the (scalar) estimator $\tilde{\theta}$.

From the previous section we know that $\tilde{\theta}$ could for example be the prediction at at covariate x_0? It would be $\tilde{\theta} = f(x_0) = x_0^T\beta$, and then $\text{MSE}(\tilde{\theta})$ would be an interesting quantity.

$$\text{MSE}(\tilde{\theta}) = \text{E}[(\tilde{\theta} - \theta)^2] = \text{Var}(\tilde{\theta}) + [\text{E}(\tilde{\theta}) - \theta]^2$$

The last transition: add and subtract $\text{E}(\tilde{\theta})$.

The first term is the variance, and the second the squared bias. (There is no irredusible error since we are not considering a new observation, but we may of cause do that and add the irreducible error.)

We know that for unbiased estimators (bias equal to 0), the MSE will be the smallest for the LS-estimator. This means that if we want to try to get a lower MSE we can´t do that with an unbiased estimator!

This is a bit unusual to many of us, since we from our first course in statistics have been told about the glory of unbiased estimators!

But, if we shrink some of the regression coefficients towards 0, or set them equal to 0, then we get a *biased estimate* for the regression parameters. Biased estimates are the core of this part of the course. We may want to pay the price of a biased estimate with the gain of decreased variance, so that the MSE for might get lower than for the LS-estimate.

Preparing for shrinkage

Standarization of covariates

For shrinkage methods it is common to *standardize* the covariates, where standardize means that

- the covariates are first centered, that is $\frac{1}{N}\sum_{i=1}^N x_{ij} = 0$ for all $j = 1, \ldots, p$,
- and then scaled to unit variance, that is $\frac{1}{N}\sum_{i=1}^N x_{ij}^2 = 1$.

This is done in practice by first subtracting the mean and then dividing by the standard deviation. The standarization is only needed if the covariates are of different units or scales, because for shrinkage we will (for some of the method) penalize the optimization with the same penalty for all covariates.

Centering covariates and response

The intercept term β_0 will not be the aim for shrinkage in shrinkage methods.

To make the presentation of the shrinkage methods easier to explain and write down, HTW use the common trick to center all covariates *and* the response.

By centering the covariates and the response we may imagine moving the centroide of the data to the origin, where we do not need an intercept to capture the best linear regression hyperplane.

When both covariates and responses are centred the LS estimate for the intercept β_0 will be $\hat{\beta}_0 = 0$ (see exercise "Centering"). If interpretation is to be done for uncentered data we may calculate the estimated β_0 for uncentered data from the estimated regression coefficients and the mean of the original covariates and respons.

When covariates and responses are centred HTW remove β_0 from the regression model for the shrinkage methods.

Group discussion: Why is the LS estimate $\hat{\beta}_0 = 0$ for centered covariates and response in the multiple linear regression model?

AND: explain what is done in the analysis of the Gasoline data directly below.

Gasoline data

Consider the multiple linear regression model, with response vector $\mathbf{Y}$ of dimension $(N \times 1)$ and p covariates and intercept in $\mathbf{X}$

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon$$

where $\varepsilon \sim \mathbf{N}(\mathbf{0}, \sigma^2\mathbf{I})$.

When gasoline is pumped into the tank of a car, vapors are vented into the atmosphere. An experiment was conducted to determine whether Y, the amount of vapor, can be predicted using the following four variables based on initial conditions of the tank and the dispensed gasoline:

- x_1: `TankTemp` tank temperature (^{o}F)
- x_2: `GasTemp` gasoline temperature (^{o}F)
- x_3: `TankPres` vapor pressure in tank (psi)
- x_4: `GasPres` vapor pressure of gasoline (psi)

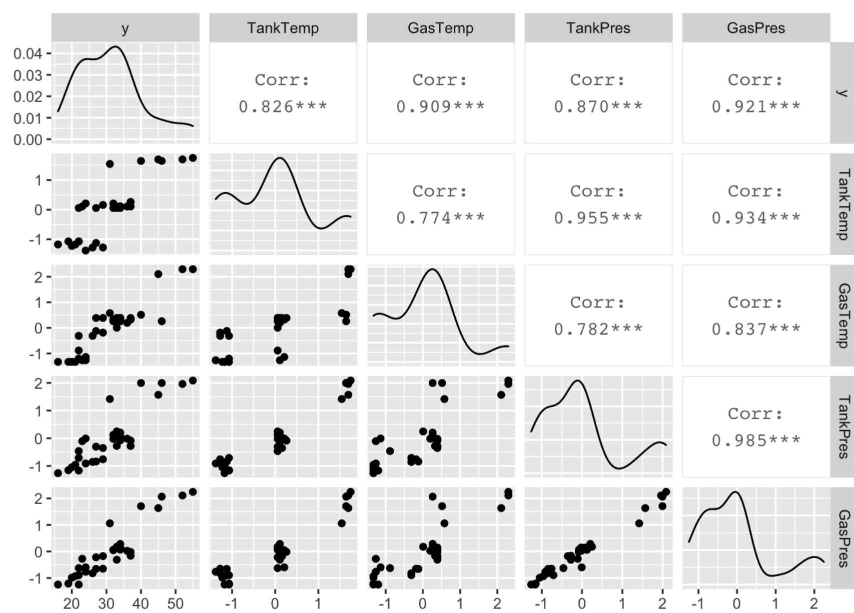
The data set is called `sniffer.dat`.

We start by standardizing the covariates (make the mean 0 and the variance 1) and then make scatter plots of the response and the covariates. Does this point to a MLR model?

```
ds <- read.table("./sniffer.dat",header=TRUE)
x <- apply(ds[,-5],2,scale)
y <- ds[,5]
print(dim(x))
```

```
## [1] 32  4
```

```
ggpairs(data.frame(y,x))
```



Calculate the estimated covariance matrix of the standardized covariates. Do you see a potential problem here?

```
cov(x)
```

```
##           TankTemp   GasTemp  TankPres   GasPres
## TankTemp 1.0000000 0.7742909 0.9554116 0.9337690
## GasTemp  0.7742909 1.0000000 0.7815286 0.8374639
## TankPres 0.9554116 0.7815286 1.0000000 0.9850748
## GasPres  0.9337690 0.8374639 0.9850748 1.0000000
```

We have fitted a MLR with all four covariates. Explain what you see.

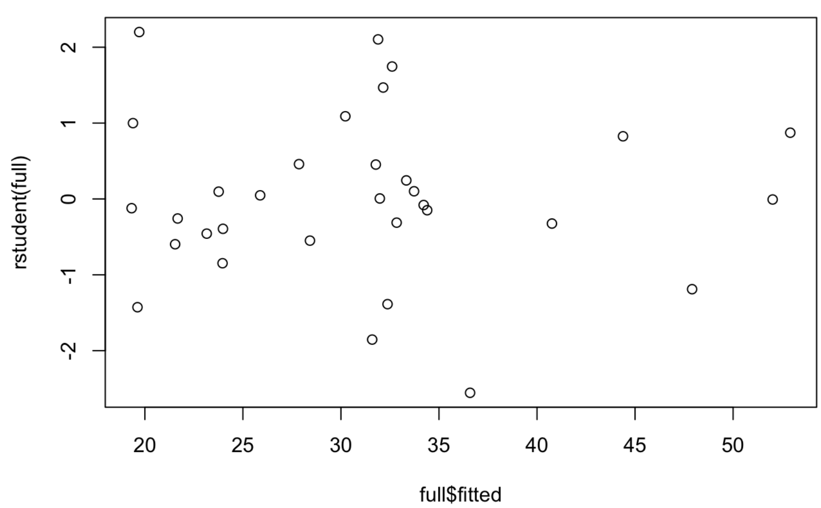
```
full <- lm(y~x)
summary(full)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.586 -1.221 -0.118  1.320  5.106
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  31.1250     0.4826  64.494  < 2e-16 ***
## xTankTemp    -0.5582     1.7677  -0.316  0.75461
## xGasTemp      3.3953     1.0654   3.187  0.00362 **
## xTankPres    -6.2737     4.1403  -1.515  0.14132
## xGasPres     12.4904     3.8587   3.237  0.00319 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.73 on 27 degrees of freedom
## Multiple R-squared:  0.9261, Adjusted R-squared:  0.9151
## F-statistic: 84.54 on 4 and 27 DF,  p-value: 7.249e-15
```

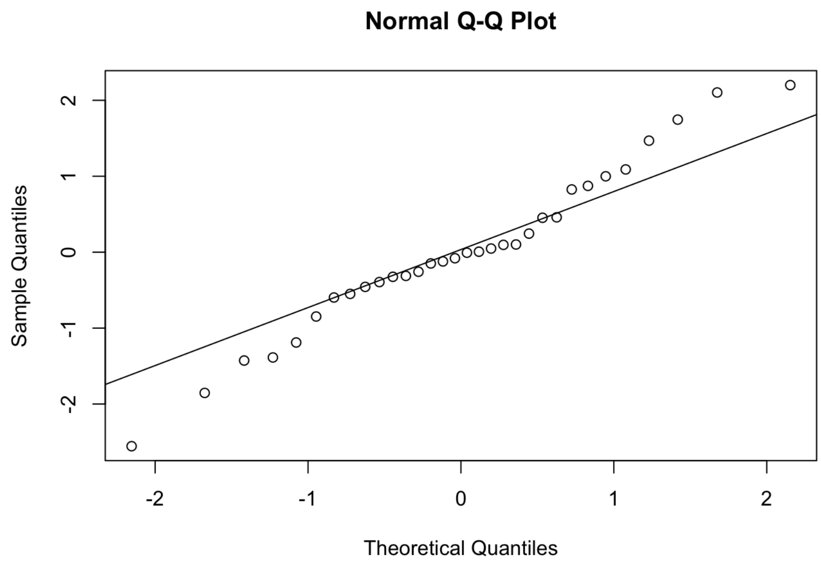
```
confint(full)
```

```
##                  2.5 %    97.5 %
## (Intercept)  30.134788 32.115212
## xTankTemp    -4.185204  3.068844
## xGasTemp      1.209363  5.581255
## xTankPres   -14.768913  2.221418
## xGasPres      4.573047 20.407838
```

```
plot(full$fitted,rstudent(full))
```




```
qqnorm(rstudent(full))
qqline(rstudent(full))
```




```
print(ad.test(rstudent(full)))
```

```
##
##  Anderson-Darling normality test
##
## data:  rstudent(full)
## A = 0.3588, p-value = 0.43
```

Perform best subset selection using Mallows C_p (equivalent to AIC) to choose the best model.

```
bests <- regsubsets(x,y)
sumbests <- summary(bests)
print(sumbests)
```

```
## Subset selection object
## 4 Variables  (and intercept)
##          Forced in Forced out
## TankTemp     FALSE      FALSE
## GasTemp      FALSE      FALSE
## TankPres     FALSE      FALSE
## GasPres      FALSE      FALSE
## 1 subsets of each size up to 4
## Selection Algorithm: exhaustive
##          TankTemp GasTemp TankPres GasPres
## 1  ( 1 ) " "      " "     " "      "*"
## 2  ( 1 ) " "      "*"     " "      "*"
## 3  ( 1 ) " "      "*"     "*"      "*"
## 4  ( 1 ) "*"      "*"     "*"      "*"
```

```
which.max(sumbests$adjr2)
```

```
## [1] 3
```

```
which.min(sumbests$cp)
```

```
## [1] 3
```

Ridge regression

(ELS 3.4.1)

Ridge regression is also called “Tikhonov regularization”.

We consider the classical linear model set-up, as for the LS estimation, but now we look at shrinking the coefficients towards 0 to construct biased estimators - and then "hope" that this also has made the variances decrease.

The ridge solution is dependent on the scaling of the covariates, and usually we work with standardized covariates and also with centered response.

We will not shrink the intercept β_0, because then the this will depend on the origin of the response.

Minimization problem

Budget version

We want to constrain the size of the estimated regression parameters, so we give the sum of squared regression coefficients a budget t.

Minimize the squared error loss

$$\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2$$

subject to $\sum_{j=1}^{p} \beta_j^2 \leq t$. The solution is called $\hat{\beta}_{\text{ridge}}$.

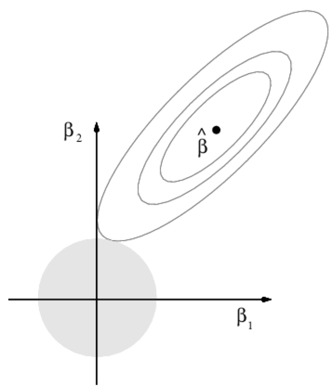


Figure from An Introduction to Statistical Learning, with applications in R (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani.

Penalty version

$$\hat{\beta}_{\text{ridge}} = \sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

where $\lambda \leq 0$ is a complexity (regularization, penalty) parameter controlling the amount of shrinkage.

- The larger λ the greater the amount of shrinkage
- The shrinkage is towards 0

This version of the problem is also called the Lagrangian form.

The budget and penalty minimization problems are equivalent ways to write the ridge regression and there is a one-to-one correspondence between the budget t and the penalty λ.

Parameter estimation

As explained, centred covariates and responses are used - and the intercept term is removed from the model. Then $\mathbf{X}$ does not include a column with 1s and has dimension $N \times p$.

Penalty criterion to minimize

$$(\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta) + \lambda\beta^T\beta$$

This can be rewritten as

$$\mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{X}\beta + \beta^T(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})\beta$$

Proceeding along the lines as done with the LS estimation, we get

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

Observe that the solution adds a positive constant λ to the diagonal of $\mathbf{X}^T\mathbf{X}$, so that even if $\mathbf{X}^T\mathbf{X}$ does not have full rank then the problem is non-singular and we can invert $(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})$.

When ridge regression was introduced in statistics in the 1970s this (avoiding non-singuarlity) was the motivation.

When $N < p$ then the design matrix will have rank less than the number of covariates, and the LS estimate does not exist. The case when two or more covariates are perfectly linearly dependent is called *super-collinearity*.

Orthogonal covariates

We study the special case with orthogonal covariates for LS and ridge.

BY HAND - SEE notes 2014.

In the special case that the columns of the design matrix are orthogonal the ridge estimates are

$$\hat{\beta}_{\text{ridge}} = \frac{1}{1+\lambda}\hat{\beta}$$

that is, a scaled version of the LS estimates.

Gasoline continued

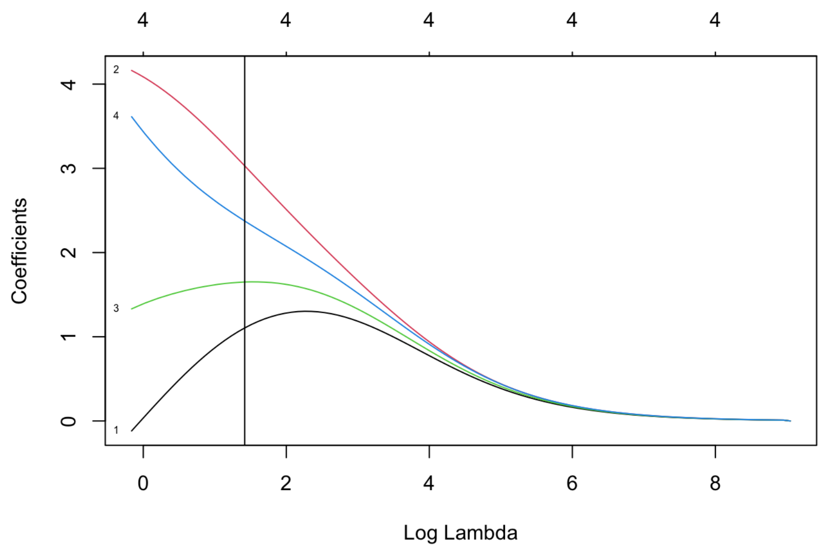
```
##   [1] 90.72273 90.36885 90.13894 90.02665 89.95213 89.87049 89.78103 89.6
8303
##   [9] 89.57571 89.45818 89.32951 89.18867 89.03457 88.86598 88.68162 88.4
8007
##  [17] 88.25982 88.01922 87.75652 87.46982 87.15711 86.81622 86.44485 86.0
4055
##  [25] 85.60075 85.12272 84.60361 84.04045 83.43017 82.76957 82.05544 81.2
8450
##  [33] 80.45347 79.55913 78.59834 77.56814 76.46577 75.28878 74.03513 72.7
0323
##  [41] 71.29210 69.80140 68.23161 66.58407 64.86110 63.06604 61.20337 59.2
7869
##  [49] 57.29877 55.27150 53.20584 51.11170 48.99983 46.88163 44.76892 42.6
7369
##  [57] 40.60788 38.58307 36.61024 34.69949 32.85985 31.09904 29.42329 27.8
3723
##  [65] 26.34462 24.94686 23.64482 22.43734 21.32215 20.29603 19.35497 18.4
9429
##  [73] 17.70885 16.99346 16.34263 15.75089 15.21285 14.72337 14.27782 13.8
7181
##  [81] 13.50149 13.16276 12.85254 12.56808 12.30659 12.06611 11.84460 11.6
4042
##  [89] 11.45203 11.27764 11.11647 10.96799 10.83017 10.70318 10.58589 10.4
7712
##  [97] 10.37671 10.28420 10.19812 10.13085
```

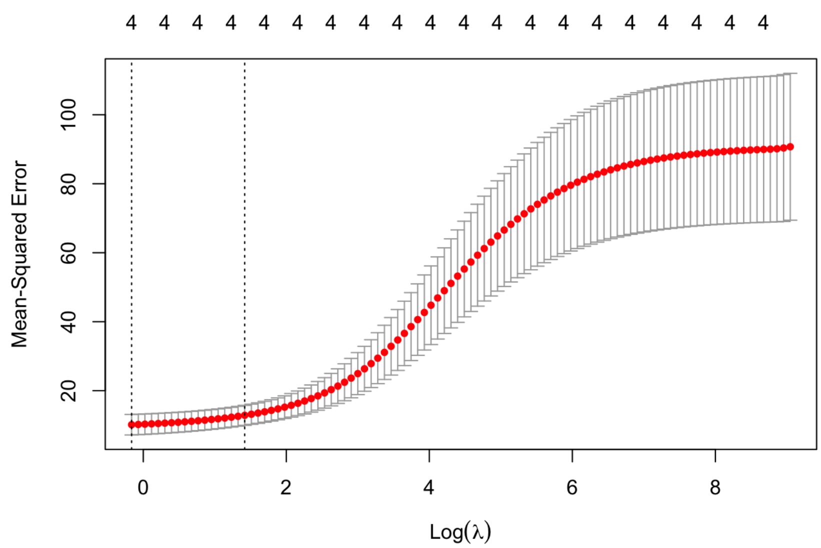
```
##   [1] 8496.6148886 7741.7990402 7054.0389514 6427.3775733 5856.3870647
##   [6] 5336.1217793 4862.0754278 4430.1420476 4036.5804384 3677.9817579
##  [11] 3351.2399957 3053.5250712 2782.2583200 2535.0901593 2309.8797367
##  [16] 2104.6763874 1917.7027380 1747.3393123 1592.1105038 1450.6717948
##  [21] 1321.7981109 1204.3732099 1097.3800134  999.8917976  911.0641662
##  [26]  830.1277367  756.3814766  689.1866310  627.9611902  572.1748488
##  [31]  521.3444123  475.0296116  432.8292902  394.3779290  359.3424808
##  [36]  327.4194852  298.3324406  271.8294088  247.6808334  225.6775508
##  [41]  205.6289792  187.3614674  170.7167911  155.5507819  141.7320791
##  [46]  129.1409919  117.6684621  107.2151202   97.6904245   89.0118764
##  [51]   81.1043066   73.8992236   67.3342202   61.3524337   55.9020526
##  [56]   50.9358683   46.4108662   42.2878527   38.5311164   35.1081183
##  [61]   31.9892098   29.1473766   26.5580040   24.1986641   22.0489215
##  [66]   20.0901561   18.3054020   16.6792005   15.1974663   13.8473653
##  [71]   12.6172035   11.4963259   10.4750240    9.5444518    8.6965490
##  [76]    7.9239715    7.2200277    6.5786204    5.9941939    5.4616862
##  [81]    4.9764851    4.5343878    4.1315653    3.7645285    3.4300981
##  [86]    3.1253777    2.8477277    2.5947434    2.3642336    2.1542016
##  [91]    1.9628283    1.7884560    1.6295745    1.4848076    1.3529014
##  [96]    1.2327134    1.1232025    1.0234203    0.9325024    0.8496615
```

```
## [1] 0.8496615
```

```
## [1] 100
```

```
## [1] 4.131565
```






```
## 5 x 1 sparse Matrix of class "dgCMatrix"
##                     1
## (Intercept) 31.125000
## TankTemp     1.104697
## GasTemp      3.029515
## TankPres     1.650469
## GasPres      2.374361
```

```
## (Intercept)   xTankTemp    xGasTemp   xTankPres    xGasPres
##  31.1250000  -0.5581796   3.3953090  -6.2737478  12.4904423
```

Properties of the ridge estimator

Mean

Derive the mean of the ridge estimator.

Exam problem 12 (TMA4268, 2019) (https://www.math.ntnu.no/emner/TMA4268/Exam/V2019e.pdf) with solutions (https://www.math.ntnu.no/emner/TMA4268/Exam/e2019sol.html) Alternatively: Wessel N. van Wieringen: Lecture notes on ridge regression, section 1.4 (https://arxiv.org/pdf/1509.09169.pdf) (We will refer to this note as WNvW below.)

What happens if:

- $\lambda \to 0$
- $\lambda \to \infty$

Covariance

Derive the covariance of the ridge estimator.

Same resources as above.

What happens if:

- $\lambda \to 0$
- $\lambda \to \infty$

(in our centered model without intercept)

We may also prove that the variance of the ridge estimator is smaller or equal the variance of the LS estimator. See exercise "Variance of ridge compared to LS", where we need to look at differences of covariance matrices and check for semi-definite matrix.

Insight based on SVD

Singular value decomposition (SVD)

(should be known from other courses?)

<et $\mathbf{X}$ be a $N \times p$ matrix.

SVD is a decomposition of a matrix $\mathbf{X}$ into a product of three matrices

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T.$$

$\mathbf{D}$ is an $(N \times p)$-dimensional block matrix. Its upper left block is a $(\text{rank}(\mathbf{X}) \times \text{rank}(\mathbf{X}))$-dimensional digonal matrix with the singular values on the diagonal. The remaining blocks, zero if $p = N$. The singular values are equal $\sqrt{\text{eigenvalues}(\mathbf{X}\mathbf{X}^T)} = \sqrt{\text{eigenvalues}(\mathbf{X}^T\mathbf{X})}$.

$\mathbf{U}$ is an $(n \times n)$-dimensional matrix with columns containing the left singular vectors (denoted $\mathbf{u}_i$), that is, the eigenvectors of $\mathbf{X}\mathbf{X}^T$

$\mathbf{V}$ is a $(p \times p)$-dimensional matrix with columns containing the right singular vectors (denoted $\mathbf{v}_i$), that is, the eigenvectors of $\mathbf{X}^T\mathbf{X}$.

The columns of $\mathbf{U}$ and $\mathbf{V}$ are orthogonal: $\mathbf{U}^\top\mathbf{U} = \mathbf{I}_N = \mathbf{U}\mathbf{U}^T$ and $\mathbf{V}^T\mathbf{V} = \mathbf{I}_p = \mathbf{V}\mathbf{V}^T$.

Following the derivation of WNvW page 11-12:

- If $n > p$ and the rank of $\mathbf{X}$ is p, then the LS estimator $\hat{\beta}$ can be written

$$\hat{\beta}_{\text{LS}} = \mathbf{V}(\mathbf{D}^T\mathbf{D})^{-1}\mathbf{D}^T\mathbf{U}^T\mathbf{Y}$$

- The ridge estimator $\tilde{\beta}$

$$\tilde{\beta}_{\text{ridge}} = \mathbf{V}(\mathbf{D}^T\mathbf{D} + \lambda\mathbf{I})^{-1}\mathbf{D}^T\mathbf{U}^T\mathbf{Y}$$

- The principal component regression based on the first k principal components

$$\hat{\beta}_{\text{PCR}} = \mathbf{V}_k(\mathbf{I}_{kp}\mathbf{D}^T\mathbf{D}\mathbf{I}_{pk})^{-1}\mathbf{I}_{kp}\mathbf{D}^T\mathbf{U}^T\mathbf{Y}$$

here $\mathbf{V}_k$ contains the first k right singular vectors as columns, and $\mathbf{I}_{kp}$ is obtained by $\mathbf{I}_p$ by removing the last $p - k$ columns.

Group discussion: What can we conclude from this about what the λ does?

Hint: the estimated covariance matrix for centred covariates is $\frac{1}{N}\mathbf{X}^T\mathbf{X}$. The small singular values d_j correspond to directions in the column space of $\mathbf{X}$ with small variance.

The ridge penalties shrinks the singular values. Principal components thresholds the singular values of $\mathbf{X}$, while ridge regression shrinks the singular values.

Alternatively, it is possible to consider the prediction

$$\hat{y}_{\text{LS}} = \mathbf{X}\hat{\beta}_{LS} = \cdots = \mathbf{U}\mathbf{U}^T\mathbf{y}$$

$$\hat{y}_{\text{ridge}} = \mathbf{X}\hat{\beta}_{ridge} = \cdots = \mathbf{U}\mathbf{D}^2(\mathbf{D}^2 + \lambda \mathbf{I}_p)^{-1}\mathbf{U}^T\mathbf{y}$$

$(_{j=1}^p) $ MISSING

+MISSING connection to PCA and explanation.

The effective degrees of freedom

In ELS Ch 7.6 we defined the effective number of parameters (here now referred to as the *effective degrees of freedom*) for a linear smoother $\hat{\mathbf{y}} = \mathbf{S}\mathbf{y}$ as

$$\text{df}(\mathbf{S}) = \text{trace}(\mathbf{S})$$

For ridge regression our linear smoother is

$$\mathbf{H}_\lambda = \mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T$$

$\text{df}(\lambda) = \text{tr}(\mathbf{H}_\lambda) = \text{tr}(\mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T) =$

MISSING result

- $\lambda = 0$ gives $\text{df}(\lambda) = p$
- $\lambda \to \infty$ gives $\text{df}(\lambda) =$

Lasso

(ELS 3.4.2)

Lasso regression is also called

We will not shrink the intercept β_0, because then the this will depend on the origin of the response.

Minimization problem

Budget version

We want to constrain the size of the estimated regression parameters, so we give the sum of squared regression coefficients a budget t.

Minimize the squared error loss

$$\sum_{i=1}^N (y_i - \beta_0 - \sum_{j=1}^p x_{ij}\beta_j)^2$$

subject to $\sum_{j=1}^p |\beta_j| \leq t$. The solution is called $\hat{\beta}_{\text{lasso}}$.

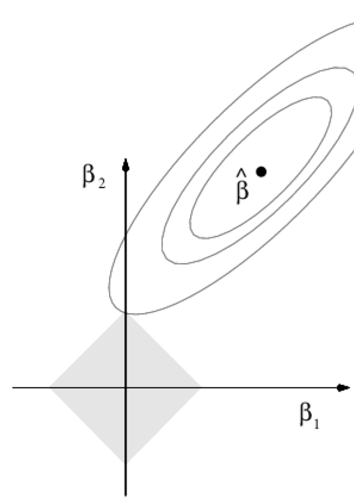


Figure from An Introduction to Statistical Learning, with applications in R (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani.

Penalty version

$$\hat{\beta}_{\text{ridge}} = \sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

where $\lambda \leq 0$ is a complexity (regularization, penalty) parameter controlling the amount of shrinkage.

- The larger λ the greater the amount of shrinkage
- The shrinkage is towards 0

This version of the problem is also called the Lagrangian form.

The budget and penalty minimization problems are equivalent ways to write the ridge regression and there is a one-to-one correspondence between the budget t and the penalty λ.

Parameter estimation

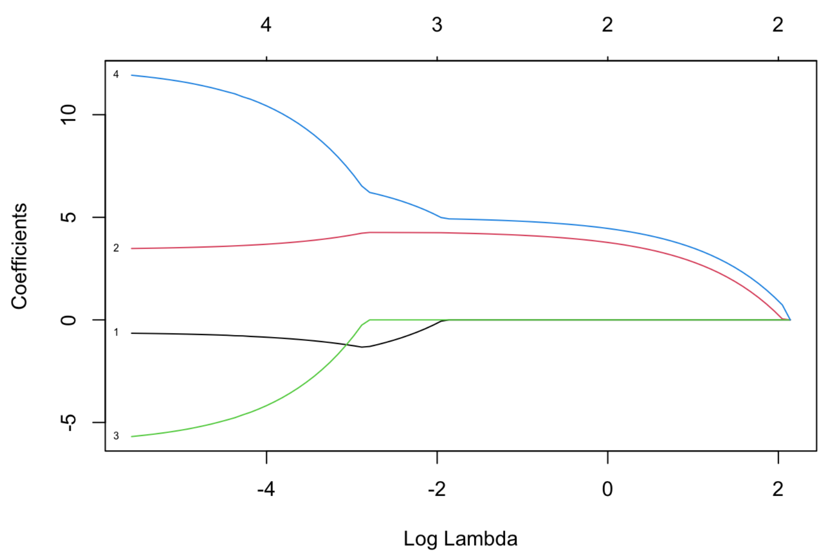
As explained, centred covariates and responses are used - and the intercept term is removed from the model. Then $\mathbf{X}$ does not include a column with 1s and has dimension $N \times p$.

In general no closed form solution.

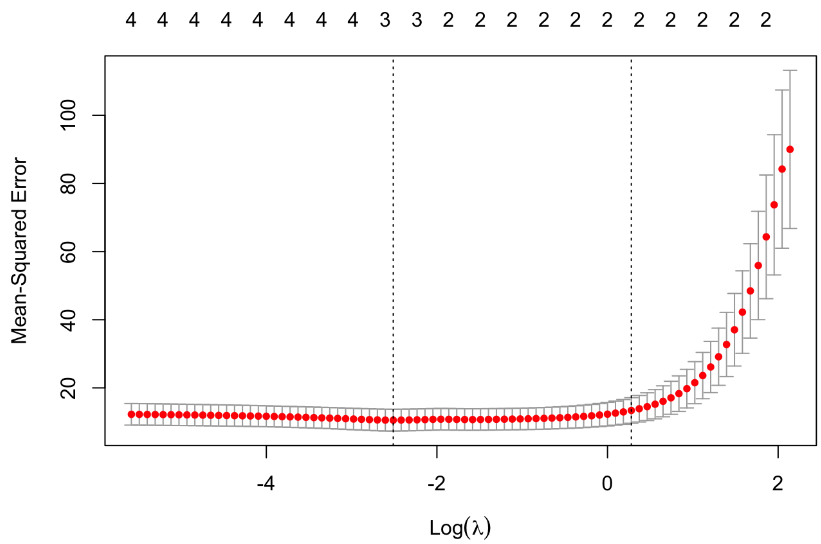
Orthogon covarates

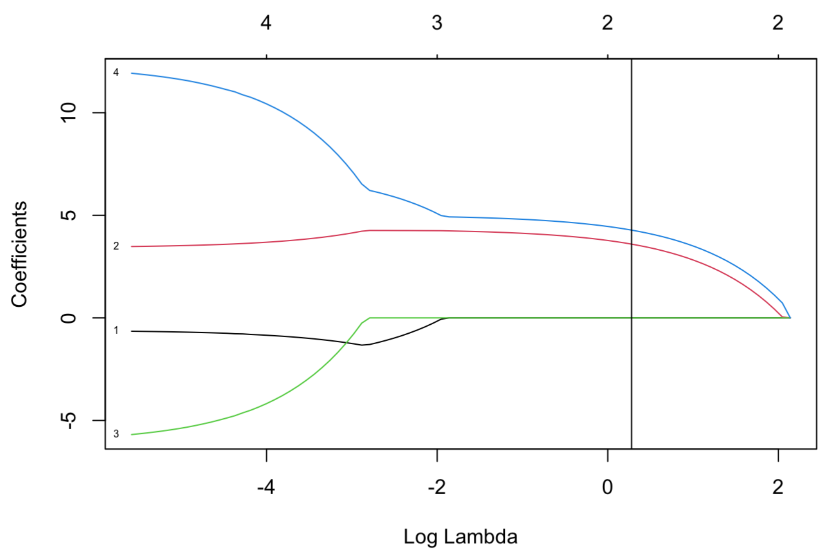
This case - explicit solutions!

Gasoline continued




```
## [1] 51
```






```
## 5 x 1 sparse Matrix of class "dgCMatrix"
##                     1
## (Intercept) 31.125000
## TankTemp     .
## GasTemp      3.594950
## TankPres     .
## GasPres      4.278981
```

Software

We will use the `glmnet` implementation for R:

- R glmnet on CRAN (https://cran.r-project.org/web/packages/glmnet/index.html) with resources (http://www.stanford.edu/~hastie/glmnet).
    - Getting started (https://glmnet.stanford.edu/articles/glmnet.html)
    - GLM with glmnet (https://glmnet.stanford.edu/articles/glmnetFamily.html)

For Python there are different options.

- Python glmnet (https://web.stanford.edu/~hastie/glmnet_python/) is recommended by Hastie et al.
- scikit-learn (https://scikit-learn.org/stable/modules/linear_model.html#ridge-regression-and-classification) (seems to mostly be for regression? is there lasso for classification here?)

Exercises

Gauss-Markov theorem

The LS is unbiased with the smallest variance among linear predictors: ELS exercise 3.3a

Variance of ridge compared to LS

Consider a classical linear model with regression parameters β. Let $\hat{\beta}$ be the LS estimator for β and let $\tilde{\beta}$ be the ridge regression estimator for β. Show that $\text{Var}(\hat{\beta}) \geq \text{Var}(\tilde{\beta})$.

Ridge regression

This problem is taken, with permission from Wessel van Wieringen, from a course in High-dimensional data analysis at Vrije University of Amsterdam.

a)

Find the ridge regression solution for the data below for a general value of λ and for the simple linear regression model $Y = \beta_0 + \beta_1 X + \varepsilon$ (only apply the ridge penalty to the slope parameter, not to the intercept). Show that when λ is chosen as 40, the ridge solution fit is $\hat{Y} = 40 + 1.75X$.

Data: $\mathbf{X}^T = (X_1, X_2, \ldots, X_8)^T = (-2, -1, -1, -1, 0, 1, 2, 2)^T$, and $\mathbf{Y}^T = (Y_1, Y_2, \ldots, Y_8)^T = (35, 40, 36, 38, 40, 43, 45, 43)^T$.

b)

The coefficients β of a linear regression model, $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$, are estimated by $\hat{\beta} = (\mathbf{X}^\mathrm{T}\mathbf{X})^{-1}\mathbf{X}^\mathrm{T}\mathbf{Y}$. The associated fitted values then given by $\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^\mathrm{T}\mathbf{X})^{-1}\mathbf{X}^\mathrm{T}\mathbf{Y} = \mathbf{H}\mathbf{Y}$, where $\mathbf{H} = \mathbf{X}(\mathbf{X}^\mathrm{T}\mathbf{X})^{-1}\mathbf{X}^\mathrm{T}$. The matrix $\mathbf{H}$ is a projection matrix and satisfies $\mathbf{H} = \mathbf{H}^2$. Hence, linear regression projects the response $\mathbf{Y}$ onto the vector space spanned by the columns of $\mathbf{X}$. Consequently, the residuals $\hat{\varepsilon}$ and $\hat{\mathbf{Y}}$ are orthogonal.

Next, consider the ridge estimator of the regression coefficients: $\hat{\beta}(\lambda) = (\mathbf{X}^\mathrm{T}\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^\mathrm{T}\mathbf{Y}$. Let $\hat{\mathbf{Y}}(\lambda) = \mathbf{X}\hat{\beta}(\lambda)$ be the vector of associated fitted values.

Show that the matrix $\mathbf{Q} = \mathbf{X}(\mathbf{X}^\mathrm{T}\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^\mathrm{T}$, associated with ridge regression, is not a projection matrix (for any $\lambda > 0$). Hint: a projection matrix is idempotent (commonly used in TMA4267).

c)

Show that the ridge fit $\hat{\mathbf{Y}}(\lambda)$ is not orthogonal to the associated ridge residuals $\hat{\varepsilon}(\lambda)$ (for any $\lambda > 0$).

Solutions to exercises

Please try yourself first, or take a small peek - and try some more - before fully reading the solutions. Report errors or improvements to Mette.Langaas@ntnu.no (mailto:Mette.Langaas@ntnu.no).

- Gauss-Markov theorem 3.3a (https://github.com/mettelang/MA8701V2021/blob/main/Part1/ELSe33a.pdf)
- Variance of ridge compared to LS: page 11-12 on note by Wessel N. van Wieringen (https://arxiv.org/pdf/1509.09169.pdf)
- Ridge regression (https://github.com/mettelang/MA8701V2021/blob/main/Part1/L2exRR1.html)

- [Lasso basics?]Noe fra ELS?
- Noe om basis kjøring av ridge og lasso for regresjon i R og python?

Resources

- Videos in statistics learning with Rob Tibshirani and Daniela Witten, made for the Introduction to statistical learning Springer textbook.
    - Ridge (https://www.youtube.com/watch?v=cSKzqb0EKS0)
    - Lasso (https://www.youtube.com/watch?v=A5I1G1MfUmA)
    - Selecting tuning parameter (https://www.youtube.com/watch?v=xMKVUstjXBE)
- Video from webinar with Trevor Hastie on glmnet from 2019 (http://youtu.be/BU2gjoLPfDc)
- Lecture notes on ridge regression: Welle N. van Wieringen (https://arxiv.org/pdf/1509.09169.pdf)