# MA8701 Advanced methods in statistical inference and learning

**Lecture 7: Shrinkage. Ridge regression**

Mette Langaas

2/2/23

## Table of contents

# 1 Shrinkage and regularization

## 1.1 Literature L7

On the reading list:

- [ESL] 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. Ebook. Chapter 3.2.2, 3.4.1.

- [HTW] Hastie, Tibshirani, Wainwrigh: "Statistical Learning with Sparsity: The Lasso and Generalizations". CRC press. Ebook. Chapter 2.1.

Strongly supporting literature

- Wessel N. van Wieringen: Lecture notes on ridge regression Chapter 1. (We will refer to this note as WNvW below.)

---

## 1.2 Why this title?

Part 2 of the course is called **Shrinkage and regularization for linear and generalized linear modESL**.

This part of the course could also have been called:

- "Penalized maximum likelihood estimation"
- "Sparse models",
- "Model selection and shrinkage"

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

---

**Central 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 overall.

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 multi-collinearity and/or more covariates than observations $N << p$. The latter may occur in medicine with genetic data, where the number of patient samples is less than the number of genetic markers studied.

---

# 2 Linear models

(ELS 3.2.1-3.2.2, HTW Ch 2.1)

Note: ELS 3.2.1 should be known from before and is therefore not on the reading list. This introduction is also partly covered in the exercises in L1.

We will only consider linear models now, and move to generalized linear models next week.

## 2.1 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).

---

## 2.2 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 $\sigma_\varepsilon^2$.

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.

---

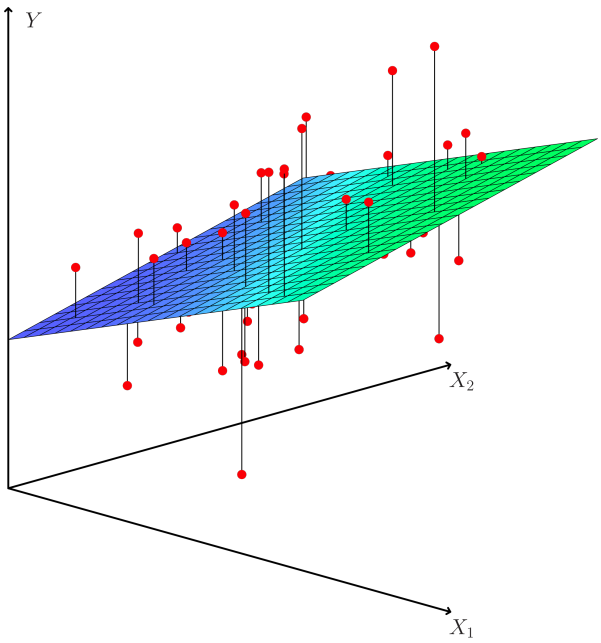



Figure 1: Figure from An Introduction to Statistical Learning, with applications in R (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani.

---

## 2.3 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.

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

Simple video introduction to categorical covariates in introductory statistical course for bachelor engineers (in Norwegian)

---

## 2.4 The classical linear model and least squares estimation

### 2.4.1 First version

For the classical linear model we assume

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

with $\mathrm{E}(\varepsilon_i) = 0$ and $\mathrm{Var}(\varepsilon_) = \sigma_\varepsilon^2$, and independence of errors $\varepsilon_j, \varepsilon_i$.

Regression parameters $\beta = (\beta_0, \beta_1, \ldots, \beta_p) \in \mathfrak{R}^{(p+1)}$.

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 $\beta$.

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

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

---

### 2.4.2 Second version

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

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

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 $\varepsilon$ $N \times 1$ random column vector

The assumptions for the classical linear model is:

1. $\mathrm{E}(\varepsilon) = \mathbf{0}$.
2. $\mathrm{Cov}(\varepsilon) = \mathrm{E}(\varepsilon\varepsilon^T) = \sigma^2\mathbf{I}$.

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 $\mathbf{X}$.

For derivation of the least squares estimator $\hat{\beta}$ see TMA4268 Module 3 and links therein.

The same results are found using likelihood theory, if we assume that $Y \sim N$. See TMA4315 GLM Module 2. Both methods are written out in these class notes from TMA4267/8.

---

The squared error loss to be minimized can be written

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

Differensiation with respect to the unknown parameter vector, and equating to zero leads to the *normal equations*.

$$\mathbf{X}^T\mathbf{X}\beta = \mathbf{X}^T\mathbf{Y}$$

To give

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

---

## 2.5 Properties of estimators

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

For the classical normal linear model:

- Least squares and maximum likelihood estimator for :

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

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

- Restricted maximum likelihood estimator for $^2$:

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

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

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

$$T_j = \frac{\hat{\beta}_{\mathrm{LS},j} - \beta_j}{\sqrt{c_{jj}}\hat{\sigma}} \sim t_{n-p-1}$$

---

## 2.6 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 $\beta$ 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 $\beta$ elements.

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

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

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

---

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 $\beta$ (not only the scalar linear combination), requires a bit more work, because we then will compare two covariance matrices.

---

### 2.6.1 Comparing variances of estimators

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

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 positive semi-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 $\tilde{\beta}$.

---

### 2.6.2 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 positive semi-definite then the variance difference will be equal or greater than 0 - by the definition of a positive semi-definite matrix.

This is also referred to as: The variance of $\tilde{\beta}$ exceeds *in a positive definite ordering sense* that of $\hat{\beta}$, and written $\text{Var}(\tilde{\beta}) \succeq \text{Var}(\hat{\beta})$. (Remark: here both Var and Cov is used as notation for the variance-covariance matrix.)

---

When is a matrix $C$ positive definite?

The matrix $C$ is positive definite if the real number $z^TCz$ is positive for every nonzero real column vector $z$.

Harville (1997)

## 2.7 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 irreducible 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.

---

## 2.8 Preparing for shrinkage

### 2.8.1 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.

---

### 2.8.2 Centering covariates and response

The intercept term $\beta_0$ 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 centered the LS estimate for the intercept $\beta_0$ will be $\widehat{\beta}_0 = 0$.

If interpretation is to be done for uncentered data we may calculate the estimated $\beta_0$ for uncentered data from the estimated regression coefficients and the mean of the original covariates and response.

When covariates and responses are centered HTW remove $\beta_0$ from the regression model for the shrinkage methods. We will also do that.

---

## 2.9 Group discussion

1) Why is the LS estimate equal to $\widehat{\beta}_0 = 0$ for centered covariates and centered response in the multiple linear regression model?

2) Explain what is done in the analysis of the Gasoline data directly below.

Choose yourself if you want to focus mainly on 1 or 2.

One version of a solution to 1 (another version in class notes L2 in 2021)

---

## 2.10 Gasoline data

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

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

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

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:

- `TankTemp` tank temperature (F)
- `GasTemp` gasoline temperature (F)
- `TankPres` vapor pressure in tank (psi)
- `GasPres` vapor pressure of gasoline (psi)

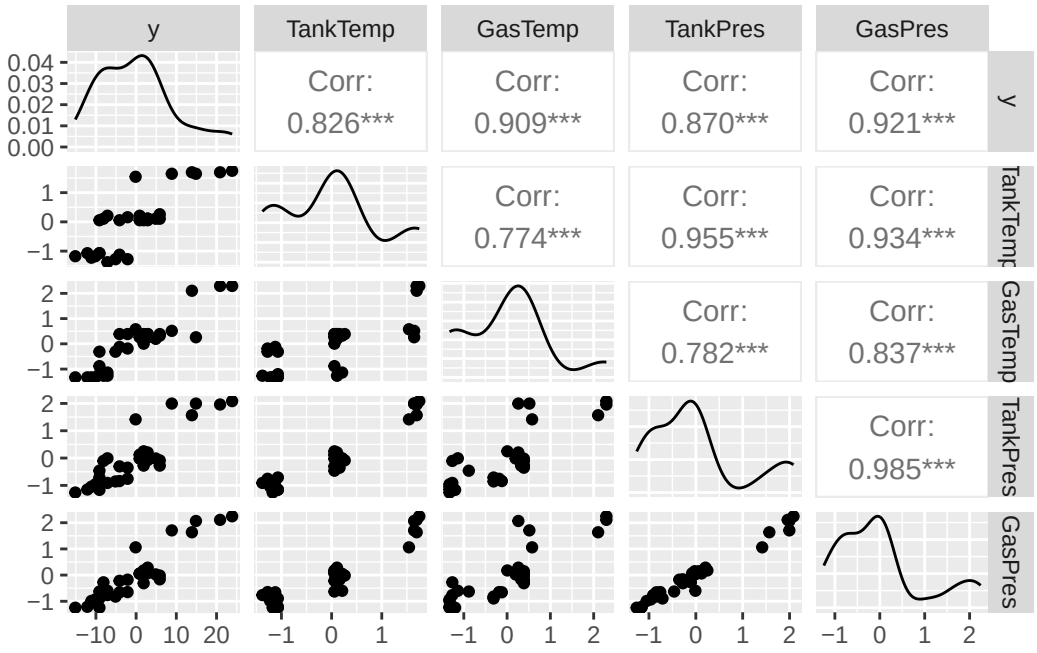
The data set is called `sniffer.dat`.

We start by standardizing the covariates (make the mean 0 and the variance 1), we also center the response. From the scatter plots of the response and the covariates - would you think an MLR is suitable?

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

```
[1] 32  4
```

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




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

```
cov(dss)
```

```
                y  TankTemp   GasTemp  TankPres   GasPres
y        87.790323 7.7399536 8.5202970 8.1505120 8.6325694
TankTemp  7.739954 1.0000000 0.7742909 0.9554116 0.9337690
GasTemp   8.520297 0.7742909 1.0000000 0.7815286 0.8374639
TankPres  8.150512 0.9554116 0.7815286 1.0000000 0.9850748
GasPres   8.632569 0.9337690 0.8374639 0.9850748 1.0000000
```

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

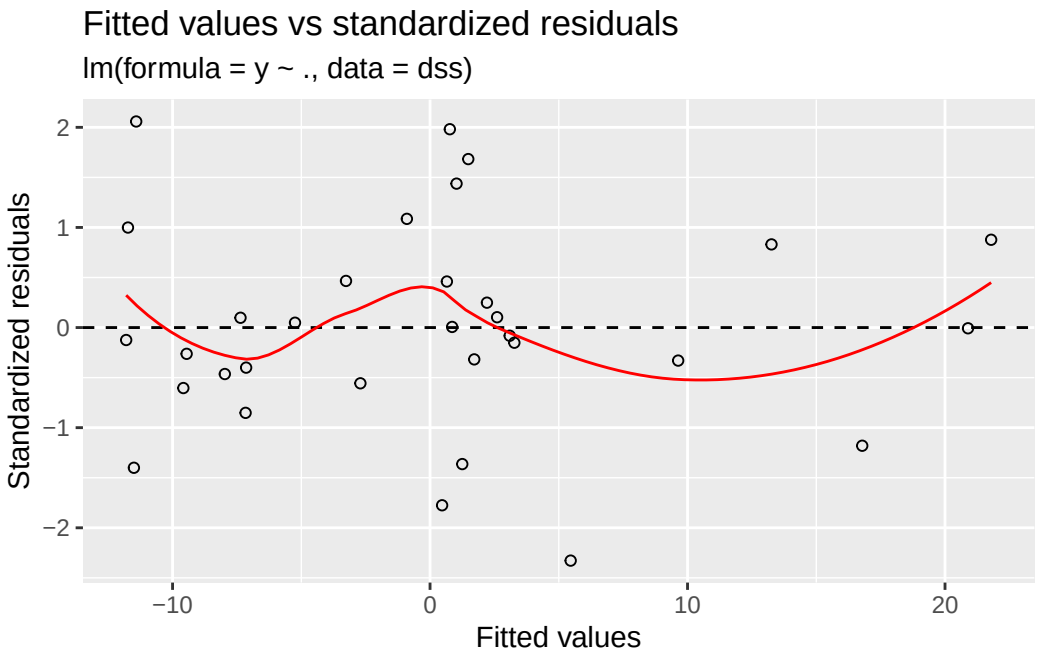
```
full <- lm(y~.,dss)
summary(full)
```

```
Call:
lm(formula = y ~ ., data = dss)

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)  3.233e-16  4.826e-01   0.000  1.00000
TankTemp    -5.582e-01  1.768e+00  -0.316  0.75461
GasTemp      3.395e+00  1.065e+00   3.187  0.00362 **
TankPres    -6.274e+00  4.140e+00  -1.515  0.14132
GasPres      1.249e+01  3.859e+00   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
```

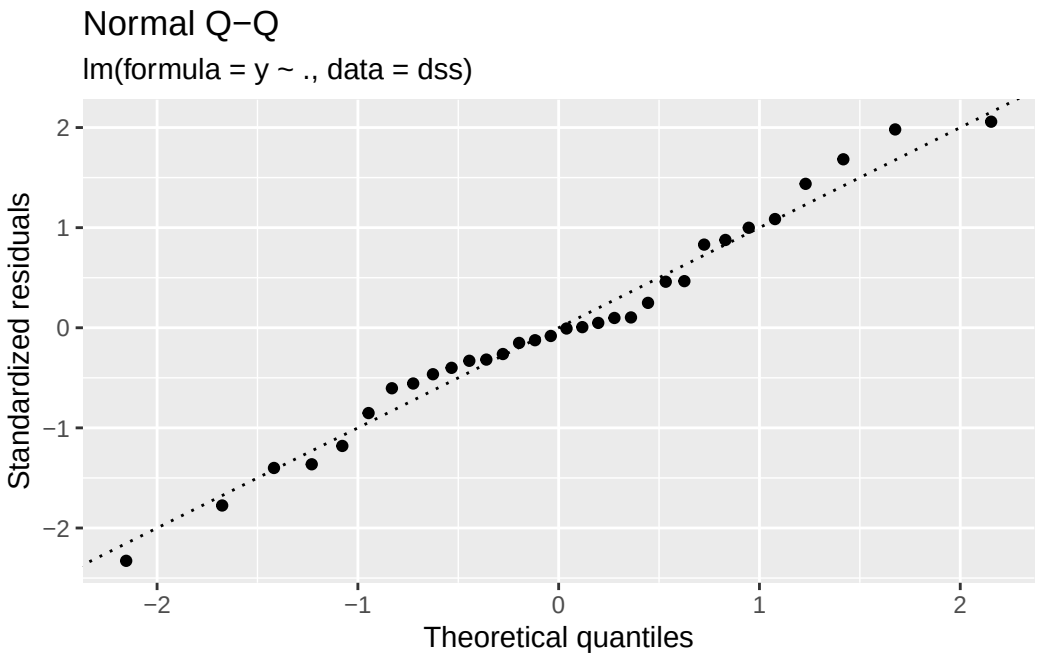
```
confint(full)
```

```
                2.5 %     97.5 %
(Intercept)  -0.9902125  0.9902125
TankTemp     -4.1852036  3.0688444
GasTemp       1.2093630  5.5812551
TankPres    -14.7689131  2.2214176
GasPres       4.5730466 20.4078380
```

```
ggplot(full, aes(.fitted, .stdresid)) + geom_point(pch = 21) + geom_hline(yintercept = 0,
    linetype = "dashed") + geom_smooth(se = FALSE, col = "red", size = 0.5,
    method = "loess") + labs(x = "Fitted values", y = "Standardized residuals",
    title = "Fitted values vs standardized residuals", subtitle = deparse(full$call))
```




```
ggplot(full, aes(sample = .stdresid)) + stat_qq(pch = 19) + geom_abline(intercept = 0,
    slope = 1, linetype = "dotted") + labs(x = "Theoretical quantiles",
    y = "Standardized residuals", title = "Normal Q-Q", subtitle = deparse(full$call))
```



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

```
[1] 3
```

Model after best subset selection.

```
red <- lm(y~GasTemp+TankPres+GasPres,data=dss)
summary(red)
```

```
Call:
lm(formula = y ~ GasTemp + TankPres + GasPres, data = dss)

Residuals:
    Min      1Q  Median      3Q     Max
-5.6198 -1.2934 -0.0496  1.4858  4.9131

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.390e-16  4.748e-01   0.000  1.00000
GasTemp      3.290e+00  9.951e-01   3.306  0.00260 **
TankPres    -7.099e+00  3.159e+00  -2.247  0.03272 *
GasPres      1.287e+01  3.607e+00   3.568  0.00132 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.686 on 28 degrees of freedom
Multiple R-squared:  0.9258,    Adjusted R-squared:  0.9178
F-statistic: 116.4 on 3 and 28 DF,  p-value: 6.427e-16
```

```r
confint(red)
```

```
                 2.5 %     97.5 %
(Intercept)  -0.9725378  0.9725378
GasTemp       1.2513019  5.3281126
TankPres    -13.5706954 -0.6270544
GasPres       5.4823283 20.2586338
```

## 3 Ridge regression

(ESL 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.

We will not shrink the intercept $\beta_0$, because then the this will depend on the origin of the response.

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

---

## 3.1 Minimization problem

### 3.1.1 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 $\hat{\beta}_{\text{ridge}}$.

---

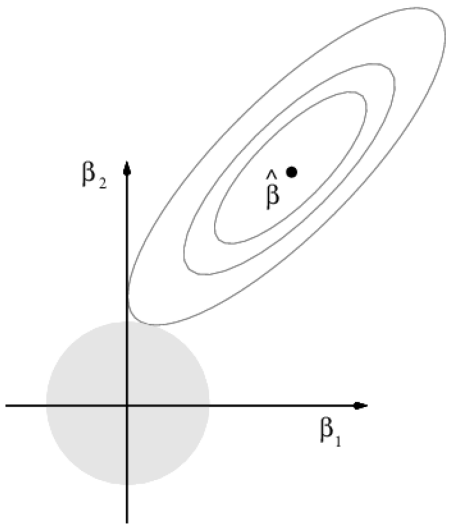



Figure 2: Figure from An Introduction to Statistical Learning, with applications in R (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani.

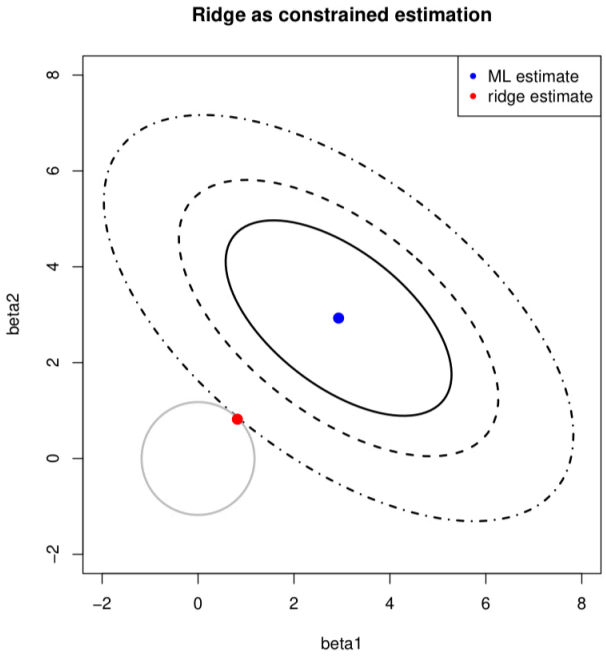



**Figure 1.4**: Top panels show examples of convex

Figure 3: Figure from Wessel N. van Wieringen: Lecture notes on ridge regression, Figure 1.4 lower left panel. CC-BY-NC-SA

### 3.1.2 Penalty version

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

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

- The larger $\lambda$ 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 $\lambda$.

---

## 3.2 Parameter estimation

As explained, centred covariates and responses are used - and the intercept term is removed from the model. Then NOW $\mathbf{X}$ 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 the (new) normal equations

$$(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})\beta = \mathbf{X}^T\mathbf{Y}$$

See "Repetition: rules for derivatives with respect to vector"

The estimator:

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

---

Observe that the solution adds a positive constant $\lambda$ 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* (accoring to WNvN).

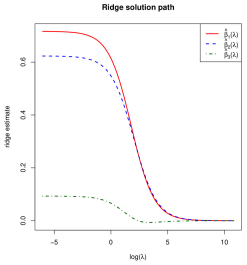
---

## 3.3 Properties of the ridge estimator

**Example 1.3** *(Super-collinearity)*
Consider the design matrix:

$$\mathbf{X} = \begin{pmatrix} 1 & -1 & 2 \\ 1 & 0 & 1 \\ 1 & 2 & -1 \\ 1 & 1 & 0 \end{pmatrix}$$

The columns of $\mathbf{X}$ are linearly dependent: the first column is the row-wise sum of the other two columns. The rank (more correct, the column rank) of a matrix is the dimension of space spanned by the column vectors. Hence, the rank of $\mathbf{X}$ is equal to the number of linearly independent columns: $\text{rank}(\mathbf{X}) = 2$. □



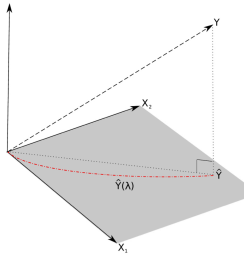




**Figure 1.1**: Left panel: the regularization path of the ridge estimator for the data of Example 1.3. Right panel: the 'maximum likelihood fit' $\widehat{Y}$ and the 'ridge fit' $\widehat{Y}(\lambda)$ (the dashed-dotted red line) to the observation $Y$ in the (hyper)plane spanned by the covariates.

Figures from Wessel N. van Wieringen: Lecture notes on ridge regression, Example 1.3 and Figure 1.1 on super-collinearity. CC-BY-NC-SA

---

### 3.3.1 Mean

Derive the mean of the ridge estimator.

What happens if:

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

---

Exam problem 12 (TMA4268, 2019) with solutions Alternatively: Wessel N. van Wieringen: Lecture notes on ridge regression, section 1.4

---

### 3.3.2 Covariance

Derive the covariance of the ridge estimator.

What happens if:

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

(in our centered model without intercept)

Same resources as above.

---

### 3.3.3 Distribution

For the normal linear model

$$\hat{\beta}(\lambda)_{\text{ridge}} \sim N\{(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^T\mathbf{X}\,\beta,$$
$$\sigma^2(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\}.$$

---

## 3.4 Is ridge "better than" LS?

1) We may 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 positive semi-definite matrix.
2) In addition it is possible to prove that given a suitable choice for $\lambda$ the ridge regression estimator may outperform the LS estimator in terms of the MSE. See WNvW Section 1.4.3 for the full derivation.
3) The optimal choice of $\lambda$ depends both the true regression parameters and the error variance. This means that the penalty parameter should be chosen in a *data-driven* fashion.

---

---

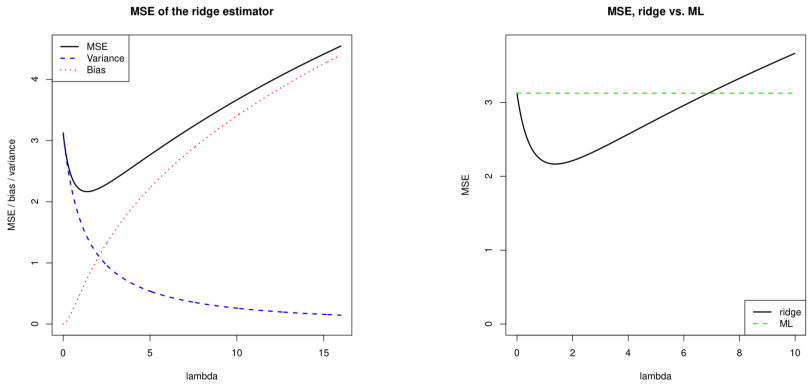



**Figure 1.3**: Left panel: mean squared error, and its 'bias' and 'variance' parts, of the ridge regression estimator (for artificial data). Right panel: mean squared error of the ridge and ML estimator of the regression coefficient vector (for the same artificial data).

Figure 4: Figures from Wessel N. van Wieringen: Lecture notes on ridge regression CC-BY-NC-SA

## 3.5 Model selection

To choose the optimal penalty parameter $\lambda$ cross-validation is the default method in use. ESL recommends to either

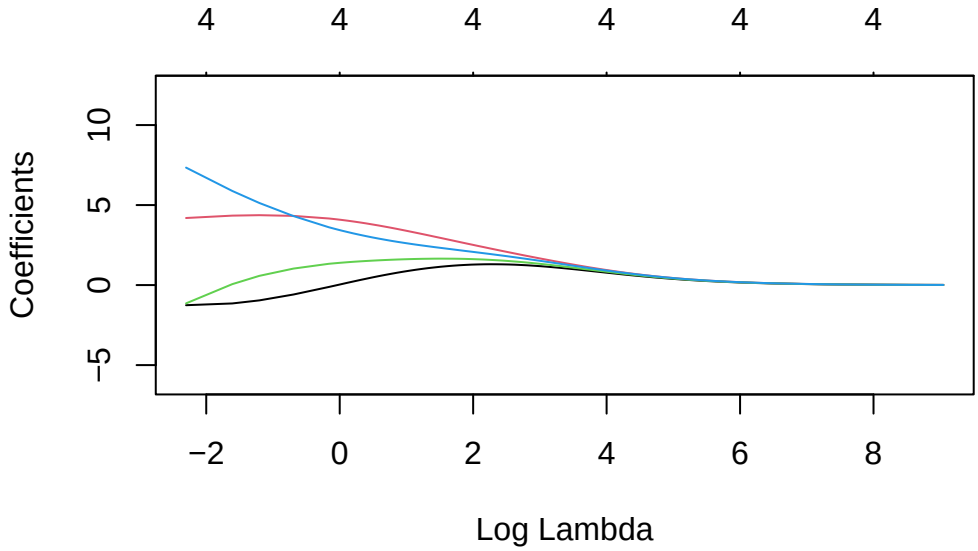
- choose the $\lambda$ corresponding to the smallest CV error
- or first find the $\lambda$ with the smallest CV-error, and then record the estimated standard error of the CV-error at this value, and then choose the largest $\lambda$ such that the CV error is still within one standard error of the minimum. We choose the largest because we want the less flexible model.

The R package `glmnet` (by Hastie et al) has default $K = 10$ fold cross-validation with the function `cv.glmnet` where `alpha=0` gives the ridge penalty.

---

### 3.5.1 Gasoline continued

```
start=glmnet(x=x,y=y,alpha=0)
autolambda=start$lambda # automatic choice of lambda had smallest lambda 0.96 - but I adde
newlambda=c(autolambda,0.5,0.3,0.2,0.1,0)
fit.ridge=glmnet(x,y,alpha=0,lambda=newlambda)
plot(fit.ridge,xvar="lambda",label=TRUE)
```




```
#plot(fit.ridge,xvar="norm",label=TRUE)
```

---

### 3.6 Group discussion

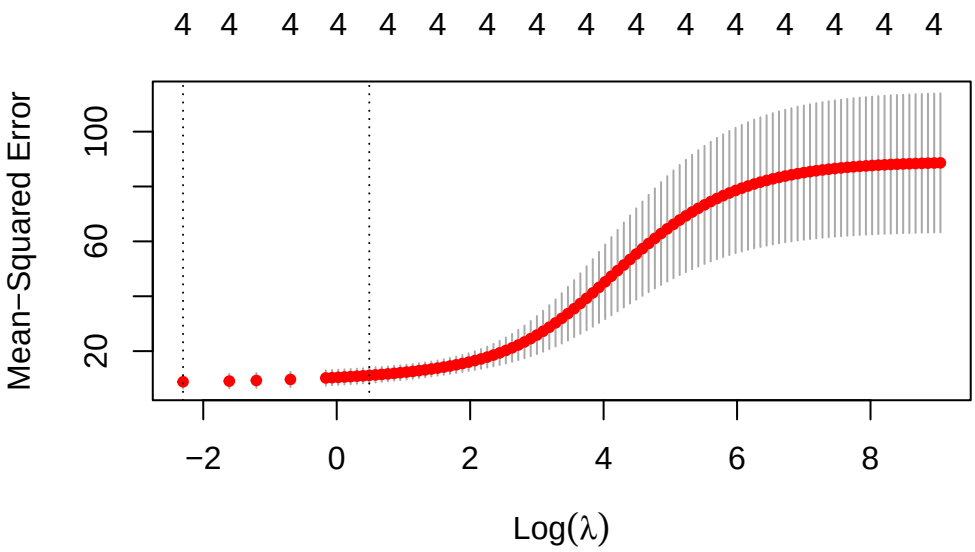
Explain what you see!

---

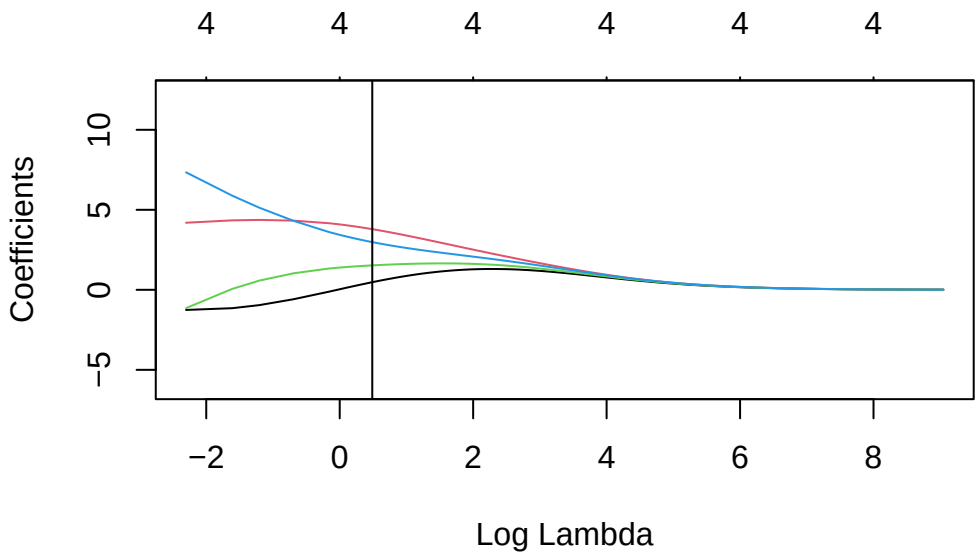
```
cv.ridge=cv.glmnet(x,y,alpha=0,lambda=newlambda)
print(cv.ridge)
```

```
Call:  cv.glmnet(x = x, y = y, lambda = newlambda, alpha = 0)

Measure: Mean-Squared Error

    Lambda Index Measure    SE Nonzero
min   0.10   104   8.838 2.426       4
1se   1.63    93  11.169 2.751       4
```

```r
#print(paste("The lamda giving the smallest CV error",cv.ridge$lambda.min))
#print(paste("The 1se method lambda",cv.ridge$lambda.1se))

plot(cv.ridge)
```




```r
# use 1se error rule default
plot(fit.ridge,xvar="lambda",label=TRUE);
abline(v=log(cv.ridge$lambda.1se));
```




```
print("Ridge 1 se method coeff")
```

```
[1] "Ridge 1 se method coeff"
```

```
coef(fit.ridge,s=cv.ridge$lambda.1se)
```

```
5 x 1 sparse Matrix of class "dgCMatrix"
                        s1
(Intercept) -2.264606e-15
TankTemp     4.770041e-01
GasTemp      3.791008e+00
TankPres     1.527068e+00
GasPres      2.981251e+00
```

```
print("LS full model coeff")
```

```
[1] "LS full model coeff"
```

```
full$coeff
```

```
  (Intercept)      TankTemp       GasTemp      TankPres       GasPres
3.232869e-16 -5.581796e-01  3.395309e+00 -6.273748e+00  1.249044e+01
```

```
print("Mallows Cp reduced model coeff")
```

```
[1] "Mallows Cp reduced model coeff"
```

```
red$coeff
```

```
  (Intercept)       GasTemp      TankPres       GasPres
8.390059e-16  3.289707e+00 -7.098875e+00  1.287048e+01
```

---

## 3.7 Insight based on SVD

(ESL 3.4.1)

### 3.7.1 Singular value decomposition (SVD)

Let $\mathbf{X}$ 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 $(p \times p)$-dimensional block matrix, with singular values on the diagonal, ordered such that $d_1 \geq d_2 \geq \ldots d_p \geq 0$. The singular values $d_j$ are equal $\sqrt{\text{eigenvalues}(\mathbf{X}\mathbf{X}^T)} = \sqrt{\text{eigenvalues}(\mathbf{X}^T\mathbf{X})}$.

---

The eigenvalues of $\mathbf{X}^T\mathbf{X}$ and $\mathbf{X}\mathbf{X}^T$ are identical, since

$$\begin{aligned}\mathbf{X}^T\mathbf{X}\mathbf{e} &= \lambda\mathbf{e} \\ \mathbf{X}\mathbf{X}^T\mathbf{X}\mathbf{e} &= \lambda\mathbf{X}\mathbf{e} \\ \mathbf{X}\mathbf{X}^T\mathbf{e}^* &= \lambda\mathbf{e}^*.\end{aligned}$$

The eigenvectors of $\mathbf{X}\mathbf{X}^T$ equals $\mathbf{X}\mathbf{e}$ where $\mathbf{e}$ are the eigenvectors of $\mathbf{X}^T\mathbf{X}$.

For a column centred matrix $\mathbf{X}$, we estimate the covariance matrix by $(N-1)\mathbf{S} = \mathbf{X}^T\mathbf{X}$.

---

$\mathbf{U}$ is an $(N \times p)$-dimensional matrix with columns containing the left singular vectors (denoted $\mathbf{u}_j$), 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 $\mathbf{V}^T\mathbf{V} = \mathbf{I}_p = \mathbf{V}\mathbf{V}^T$.

---

$$\hat{y}_{\text{LS}} = \mathbf{X}\hat{\beta}_{\text{LS}} = \mathbf{U}\mathbf{U}^T\mathbf{y} = \sum_{j=1}^{p} \mathbf{u}_j(\mathbf{u}_j^T\mathbf{y})$$

$$\hat{y}_{\text{ridge}} = \mathbf{X}\hat{\beta}_{ridge} = \mathbf{U}\mathbf{D}^2(\mathbf{D}^2 + \lambda\mathbf{I}_p)^{-1}\mathbf{U}^T\mathbf{y} = \sum_{j=1}^{p} \mathbf{u}_j(\frac{d_j^2}{d_j^2 + \lambda})(\mathbf{u}_j^T\mathbf{y})$$

---

## 3.8 Group discussion

How can you interpret this result?

---

## 3.9 The effective degrees of freedom

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

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

For ridge regression our linear smoother is

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

---

$$\text{df}(\lambda) = \text{tr}(\mathbf{H}_\lambda) = \text{tr}(\mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T) = \cdots = \sum_{j=1}^{p} \frac{d_j^2}{d_j^2+\lambda}$$

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

The $\text{df}(\lambda)$ is sometimes plotted instead of $\lambda$ on the horisontal axis when model complexity is chosen.

---

### 3.10 Finally

- When is ridge preferred to LS? When the LS estimates have high variance and many predictors are truly non-zero.
- Ridge is computationally fast.
- Ridge is not very easy to interpret, because all $p$ predictor are included in the final model.

## 4 Software


We will use the `glmnet` implementation for R:

- R glmnet on CRAN with resources.
  - Getting started
  - GLM with glmnet

For Python there are different options.

- Python glmnet is recommended by Hastie et al.
- scikit-learn (seems to mostly be for regression? is there lasso for classification here?)

## 5 Exercises

### 5.1 Gauss-Markov theorem

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

## 5.2 Variance of ridge compared to LS

Consider a classical linear model with regression parameters $\beta$. Let $\hat{\beta}$ be the LS estimator for $\beta$ and let $\tilde{\beta}$ be the ridge regression estimator for $\beta$. Show that the variance of $\tilde{\beta}$ exceeds *in a positive definite ordering sense* that of $\hat{\beta}$, and written $\text{Var}(\tilde{\beta}) \succeq \text{Var}(\hat{\beta})$.

---

## 5.3 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.

### 5.3.1 a)

Find the ridge regression solution for the data below for a general value of $\lambda$ 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 $\lambda$ is chosen as 4, the ridge solution fit is $\hat{Y} = 40 + 1.75X$.

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

### 5.3.2 b)

The coefficients $\beta$ 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).

### 5.3.3 c)

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

---

## 5.4 Orthonormal design matrix

Assume that the design matrix $\mathbf{X}$ is ortonormal, that is, $\mathbf{X}^T\mathbf{X} = \mathbf{I}_{pp} = (\mathbf{X}^T\mathbf{X})^{-1}$.

a) Derive the relationship between the least squares and the ridge regression estimator.

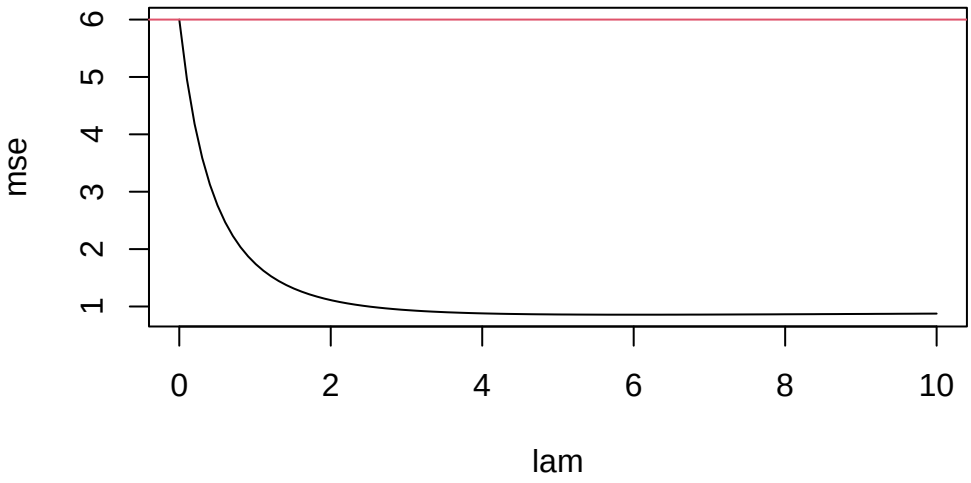
b) Derive the relationship between the covariance matrices for the two estimators.

c) Derive the MSE for each of the two estimators. Which value of the penalty parameter $\lambda$ gives the minimum value of the MSE for the ridge regression estimator?

# 6 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.

- Gauss-Markov theorem 3.3a
- Variance of ridge compared to LS: page 11-12 on note by Wessel N. van Wieringen and Mettes notes
- Ridge regression
- Orthonormal design matrix




## 7 Resources

- Videos in statistics learning with Rob Tibshirani and Daniela Witten, made for the Introduction to statistical learning Springer textbook.
  - Ridge
  - Selecting tuning parameter
- Video from webinar with Trevor Hastie on glmnet from 2019
- Lecture notes on ridge regression: Wessel N. van Wieringen

## 8 References

Harville, David A. 1997. *Matrix Algebra from a Statistician's Perspective*. Springer.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Vol. 2. Springer series in statistics New York. hastie.su.domains/ElemStatLearn.

Wieringen, Wessel N. van. 2020. "Lecture Notes on Ridge Regression." https://arxiv.org/pdf/1509.09169.pdf.