PART 2:
SHRINKAGE and REGULARIZATION
IN LINEAR MODELS (LM) and GENERALIZED LM (GLM)

MA8701 Advanced methods in statistical inference and learning

Lecture 7: Shrinkage. Ridge regression

Mette Langaas

1/29/23

Lectured 30.01.2023

Added after class in this colour.

Shrinkage and regularization

Literature L7

On the reading list:

- [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. 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.)

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.

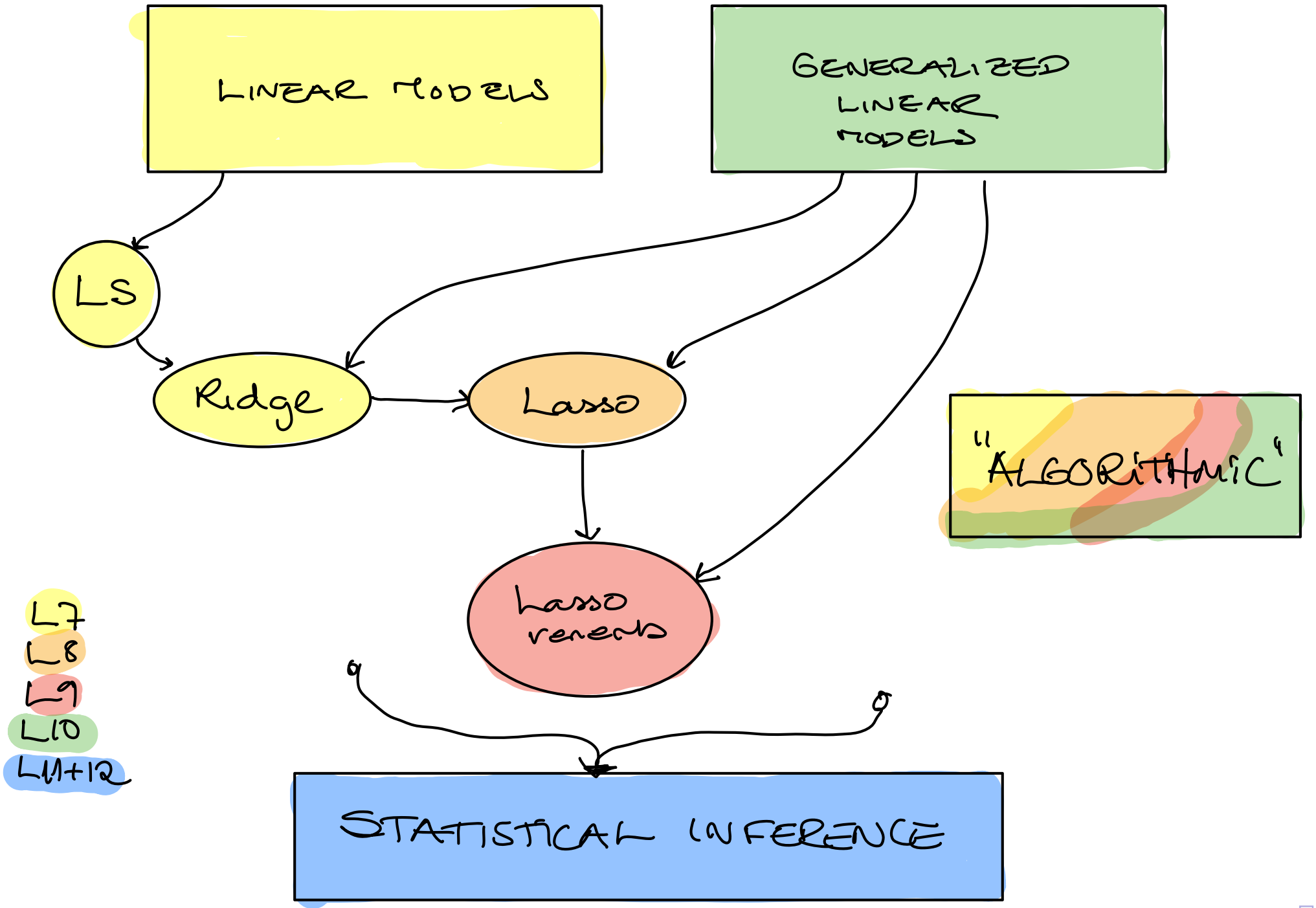
Can we use penalization (shrinkage) to produce parameter estimates with some bias but less variance, so that the prediction performance is improved?

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.


LINEAR MODELS
GENERALIZED LINEAR MODELS
LS
Ridge
Lasso
"ALGORITHMIC"
Lasso revenge
L7
L8
L9
L10
L11+12
STATISTICAL INFERENCE

Linear models

Part 1: Decision theoretic framework

$Y = f(X) + \varepsilon$, $E(\varepsilon) = 0$, $\mathrm{Var}(\varepsilon) = \sigma^2$

$f(X) = E(Y \mid X)$ optimal for squared loss

Assume $f(x_i) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}$

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

Matrix notation:

$$\underset{N\times 1}{Y} = \underset{N\times(p+1)}{X} \underset{(p+1)\times 1}{\beta} + \underset{N\times 1}{\varepsilon} \qquad E(\varepsilon) = 0,\ \mathrm{Var}(\varepsilon) = \underset{N\times N}{\sigma^2 I}$$

$\underset{\beta}{\text{minimize}} \; (Y - X\beta)^T (Y - X\beta)$

$\downarrow$

$\frac{\partial}{\partial \beta}$

$$(X^TX)\beta = X^TY \qquad \text{normal equations}$$

If (X^TX) full rank (invertible) then

$$\hat{\beta}_{LS} = (X^TX)^{-1} X^TY$$

with

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

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

$$= (X^TX)^{-1} \sigma^2$$

$Var(CY)$
$= C\,Var(Y)\,C^T$

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.*

$\hat{\beta}_{LS}$ is a linear function in Y and $E(\hat{\beta}_{LS}) = \beta$

$(X^TX)^{-1}X^TY$

If we have another $\tilde{\beta} = CY$, let $\tilde{\beta} = (X^TX)^{-1}X^TY + DY$

$$(X^TX)^{-1}X^T + D \quad (p+1)\times N$$

Then $E(\tilde{\beta}) = ((X^TX)^{-1}X^T + D)X\beta$ to β then $DX = 0$

Now to the variance: $Cov(\tilde{\beta}) = C\,\sigma^2 I\, C^T = CC^T\sigma^2$

$$CC^T = ((X^TX)^{-1}X^T + D)((X^TX)^{-1}X^T + D)^T$$

$$= (X^TX)^{-1}X^TX(X^TX)^{-1} + (X^TX)^{-1}\underbrace{X^TD^T}_{0} + \underbrace{DX}_{0}(X^TX)^{-1} + DD^T$$

$$= (X^TX)^{-1} + DD^T$$

$$Cov(\tilde{\beta}) = Cov(\hat{\beta}_{LS}) + DD^T\sigma^2$$

$$\underline{Cov(\tilde{\beta}) - Cov(\hat{\beta}_{LS}) = DD^T\sigma^2}$$

Is DD^T positive (semi) definite? Yes

$$z^T DD^T z = z^T D D^T z = \|D^T z\|_2^2 \geq 0$$

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 $\widehat{\beta}$ is smaller (or equal to) the variance of another estimator $\widetilde{\beta}$.

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

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 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 $z^T C z$ is positive for every nonzero real column vector z.

Harville (1997)

MEAN SQUARED ERROR

Let $\tilde{\theta}$ be a scalar estimator, eg. $x_0^T \tilde{\beta} = \tilde{\theta}$

Then the $MSE(\tilde{\theta}) = E\left[(\tilde{\theta} - \theta)^2\right]$ $\pm E(\tilde{\theta})$

$$= \left(E(\tilde{\theta}) - \theta\right)^2 + Var(\tilde{\theta})$$

$$\text{bias}^2$$

Vectors: $\tilde{\theta}$ $\quad tr(Var(\tilde{\theta})) + (E(\tilde{\theta}) - \theta)^T (E(\tilde{\theta}) - \theta)$

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 centered the LS estimate for the intercept β_0 will be $\widehat{\beta}_0 = 0$.

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 response.

When covariates and responses are centered HTW remove β_0 from the regression model for the shrinkage methods. We will also do that.

Group discussion

1) Why is the LS estimate equal to $\hat{\beta}_0 = 0$ 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.

Why $\hat{\beta}_0 = 0$?

2.9 Group discussion

1) Why is the LS estimate equal to $\hat{\beta}_0 = 0$ 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.

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 $\mathbf{X}$ $(N \times p + 1)$.

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

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)

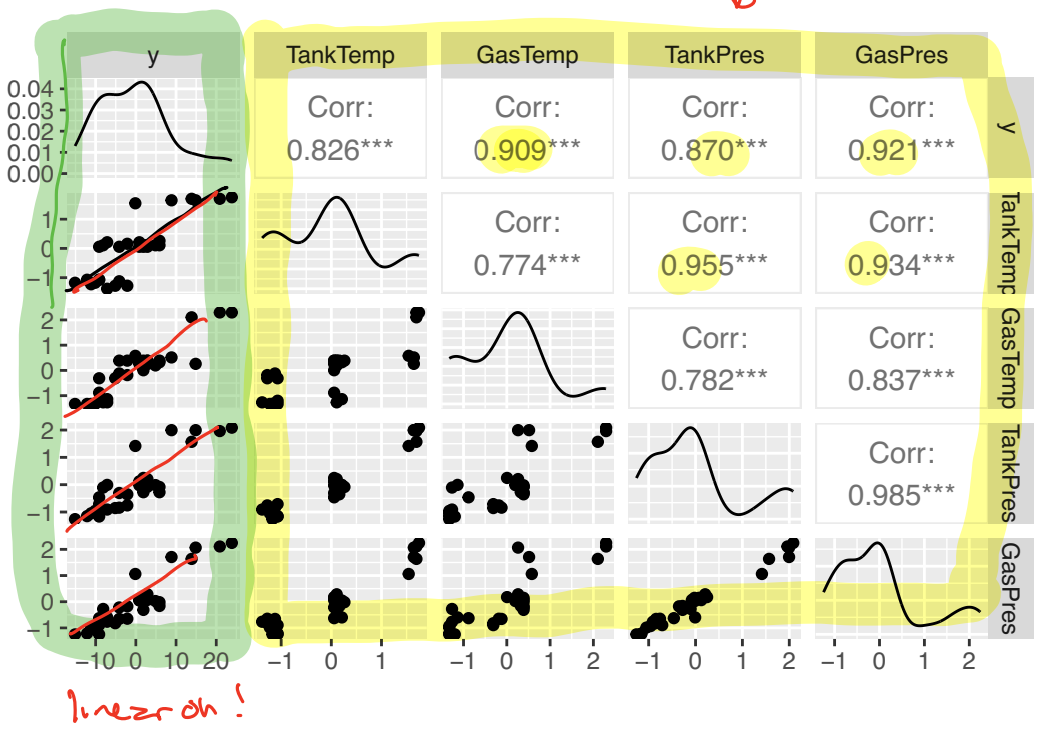
The data set is called `sniffer.dat`.

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)
```

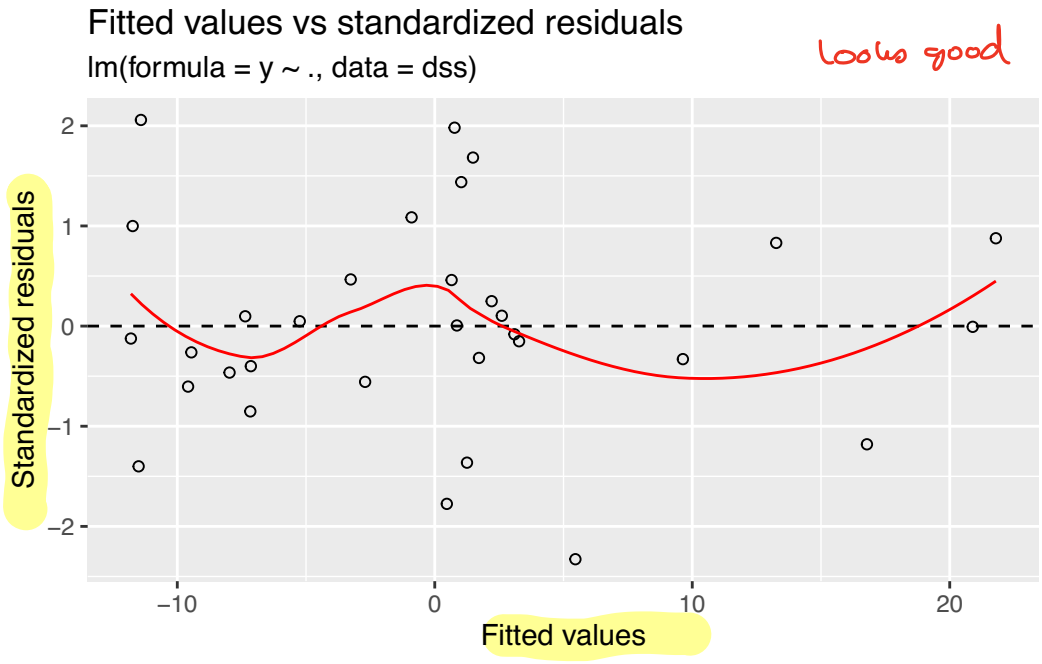
```
Call:
lm(formula = y ~ ., data = dss)

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
```

```r
confint(full)
```

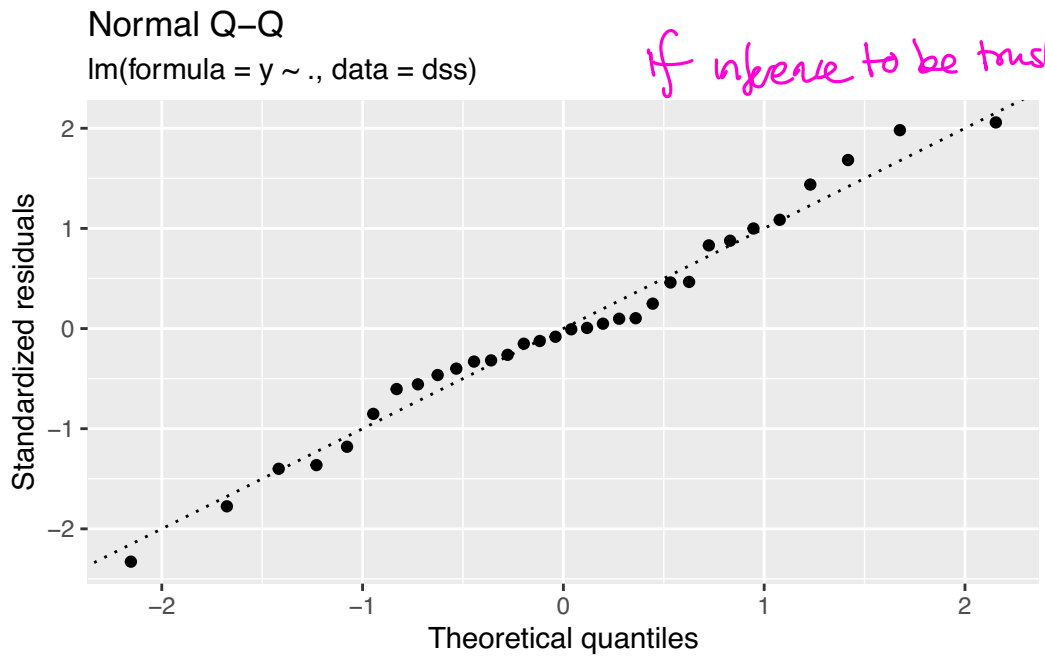
```
                2.5 %     97.5 %
(Intercept)  -0.9902125  0.9902125
TankTemp     -4.1852036  3.0688444
GasTemp       1.2093630  5.5812551
TankPres    -14.7689131  2.2214176
GasPres       4.5730466 20.4078380
```

```r
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))
```

"Most important plot"




```
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
```

H_0: residuals $\sim N$
$\Rightarrow$ not reject H_0

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

$\overline{err} + \frac{2d}{N}\hat{\sigma}_\varepsilon^2$

```
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 ) "*"      "*"      "*"      "*"
```

best model with 1 covariate

overall best

```
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
```

changes in $\hat{\beta}$, $\widehat{SE}(\hat{\beta})$!

92.6% explained

```
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

(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.

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

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.

We will not shrink the intercept β_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.

RIDGE REGRESSION

RIDGE REGRESSION

$$Y = X\beta + \varepsilon$$

$N\times 1 \quad N\times p \quad p\times 1 \quad N\times 1$

$E(\varepsilon) = 0, \; Cov(\varepsilon) = \sigma^2 I$

$(\beta_0 = 0)$

minimize $(Y - X\beta)^T (Y - X\beta)$ subject to $\beta^T\beta \le t$

wrt β

$$\sum_{j=1}^{p} \beta_j^2$$

The ridge regression estimator will <u>always</u> be on the boundry of the ridge constraint (WWW 1.5)

Alternative way of writing the problem

$$\underset{\beta}{\operatorname{argmin}} \; (Y - X\beta)^T (Y - X\beta) + \lambda \beta^T\beta$$

$$\frac{\partial (a^T\beta)}{\partial \beta} = a$$

$$\frac{\partial \beta^T A \beta}{\partial \beta} = 2A\beta$$

rewrite

$$Y^TY - 2Y^TX\beta + \beta^T(X^TX + \lambda I)\beta$$

$\frac{\partial}{\partial \beta}$ → 0

$$2(X^TX + \lambda I)\beta = 2X^TY$$

"new" normal equations

$$\hat{\beta}_{Ridge} = \hat{\beta}(\lambda) = (X^TX + \lambda I)^{-1} X^TY$$

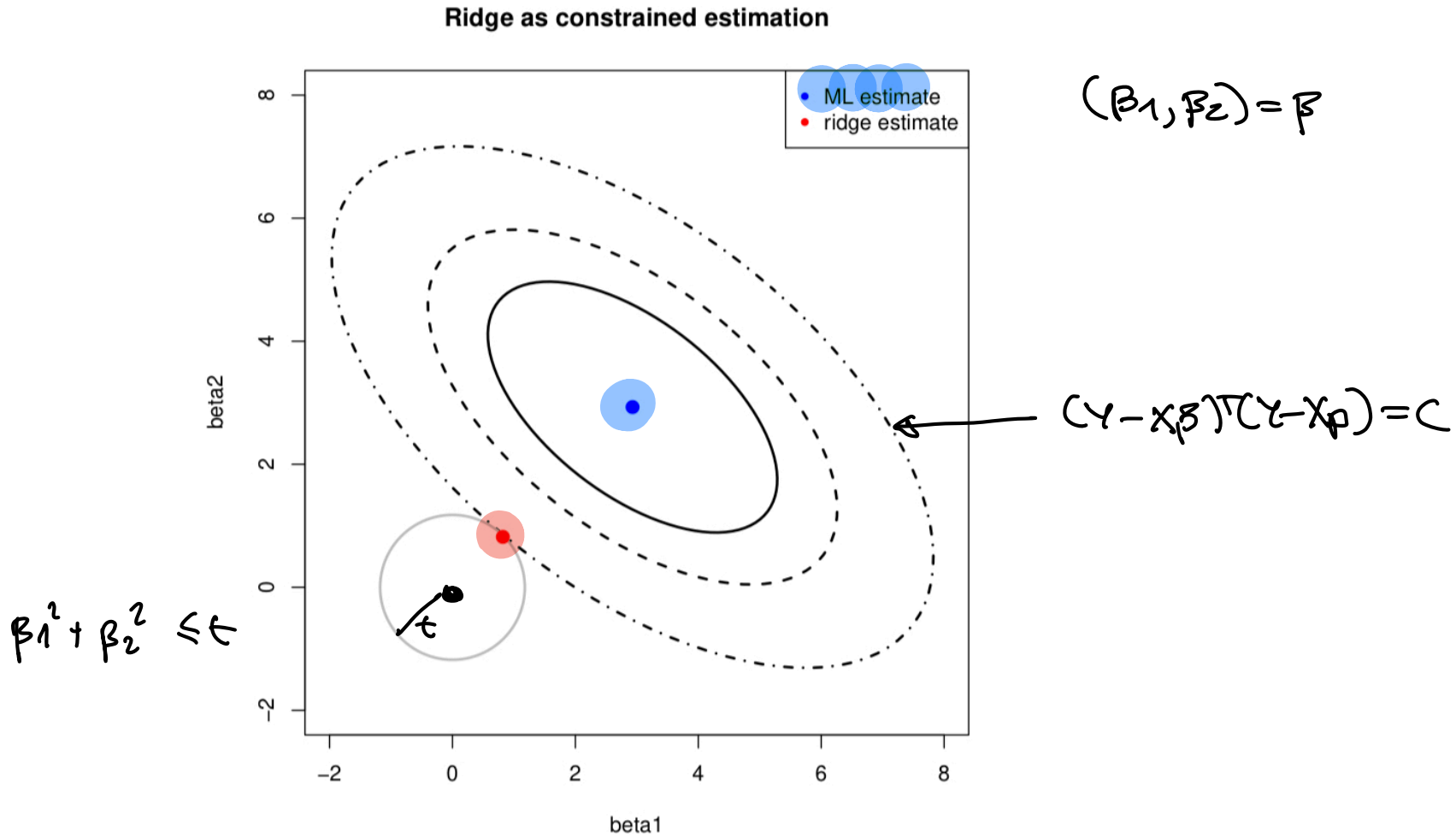


Figure 1.4: Top panels show examples of convex

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

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-singularity) 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).

PROPERTIES

$$\hat{\beta}(\lambda) = (X^TX + \lambda I)^{-1} X^TY$$

$$= \underbrace{(X^TX + \lambda I)^{-1} X^TX}_{W_\lambda} \underbrace{(X^TX)^{-1} X^TY}_{\hat{\beta}_{LS}}$$

W_λ ↑ transform from $\hat{\beta}_{LS}$ to $\hat{\beta}_{ridge}$ → use in Ex on $Var(\hat{\beta}(\lambda))$

$$\hat{Y} = X\hat{\beta}(\lambda) = \underbrace{X(X^TX + \lambda I)^{-1} X^T}_{H(\lambda)} Y = \underline{H(\lambda)Y}$$

For LS $\hat{Y}$ was projected onto the columnspace of X

$$H_{LS} = X(X^TX)^{-1}X^T \rightarrow H_{LS}^T = H_{LS} \quad \text{symmetric}$$

$$H_{LS}H_{LS} = H_{LS} \quad \text{idempotent}$$

↓

projection matrix

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: $\text{rank}(\mathbf{X}) = 2$. □

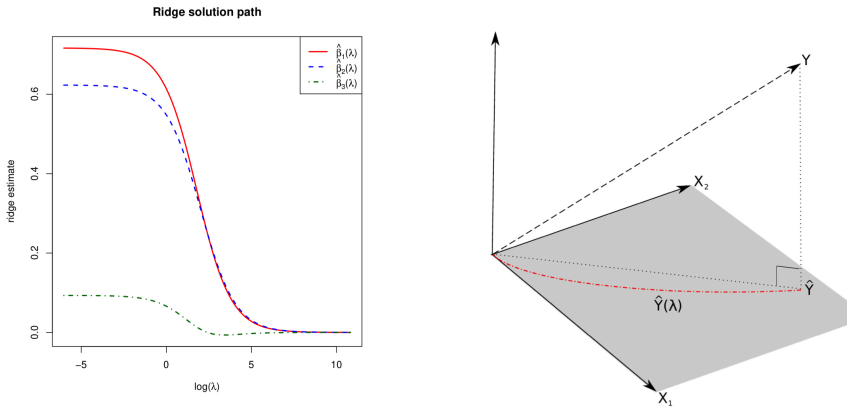


Figure 1.1: Left panel: the regularization path of the ridge estimator for the data of Example 1.3. Right panel: the 'maximum likelihood fit' $\widehat{Y}$ and the 'ridge fit' $\widehat{Y}(\lambda)$ (the dashed-dotted red line) to the observation Y in the (hyper)plane spanned by the covariates.

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

Is $H(\lambda)$ a projection matrix? No

1) $H(\lambda)^T = H(\lambda)$

2) $H(\lambda)H(\lambda) = X(X^TX+\lambda I)^{-1}X^TX(X^TX+\lambda I)^{-1}X^T \neq H(\lambda)$

Moments

Mean

Derive the mean of the ridge estimator.

What happens if:

- $\lambda \to 0$ → $\hat{\beta}_{LS}$
- $\lambda \to \infty$ $\hat{\beta}(\lambda) \to 0$ coverage does not matter

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

$E(\hat{\beta}(\lambda)) = (X^T X + \lambda I)^{-1} X^T X\beta \neq \beta$ biased unless $\lambda = 0$

$Cov(\hat{\beta}(\lambda)) = (X^TX + \lambda I)^{-1} X^T \sigma^2 I \; X (X^TX + \lambda I)^{-1}$

Covariance

Derive the covariance of the ridge estimator.

What happens if:

- $\lambda \to 0$ $Cov(\hat{\beta}(0)) = Cov(\hat{\beta}_{ls}) = (X^TX)^{-1}\sigma^2$
- $\lambda \to \infty$ $Cov(\hat{\beta}(\lambda)) \to 0$ intercept only model – no covariates will influence

(in our centered model without intercept)

Same resources as above.

DISTRIBUTION

If $\varepsilon \sim N(0, \sigma^2 I)$ then $Y \sim N(X\beta, \sigma^2 I)$

$$\hat{\beta}(\lambda) \sim N\left(E(\hat{\beta}(\lambda)), Var(\hat{\beta}(\lambda)) \right)$$

$E(\hat{\beta}(\lambda))$ not β

$H_0: \beta = 0$ vs $H_1: \beta \neq 0$

95% CI for β

How to use our classic methods when estimator is biased?

Is ridge "better than" LS?

turns out $\mathrm{Var}(\hat{\beta}(\lambda)) - \mathrm{Var}(\hat{\beta}(0))$ positive semi definite

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.

1974 theorem: there exists $\lambda > 0$ such that $\mathrm{MSE}(\hat{\beta}(\lambda)) < \mathrm{MSE}(\hat{\beta}(0))$

2) In addition it is possible to prove that given a suitable choice for λ 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 λ depends both the true regression parameters and the error variance. This means that the penalty parameter should be chosen in a *data-driven* fashion.

optimal λ dep. on β and σ^2

Ex. orthonormal X ← DO THE EXERCISE!

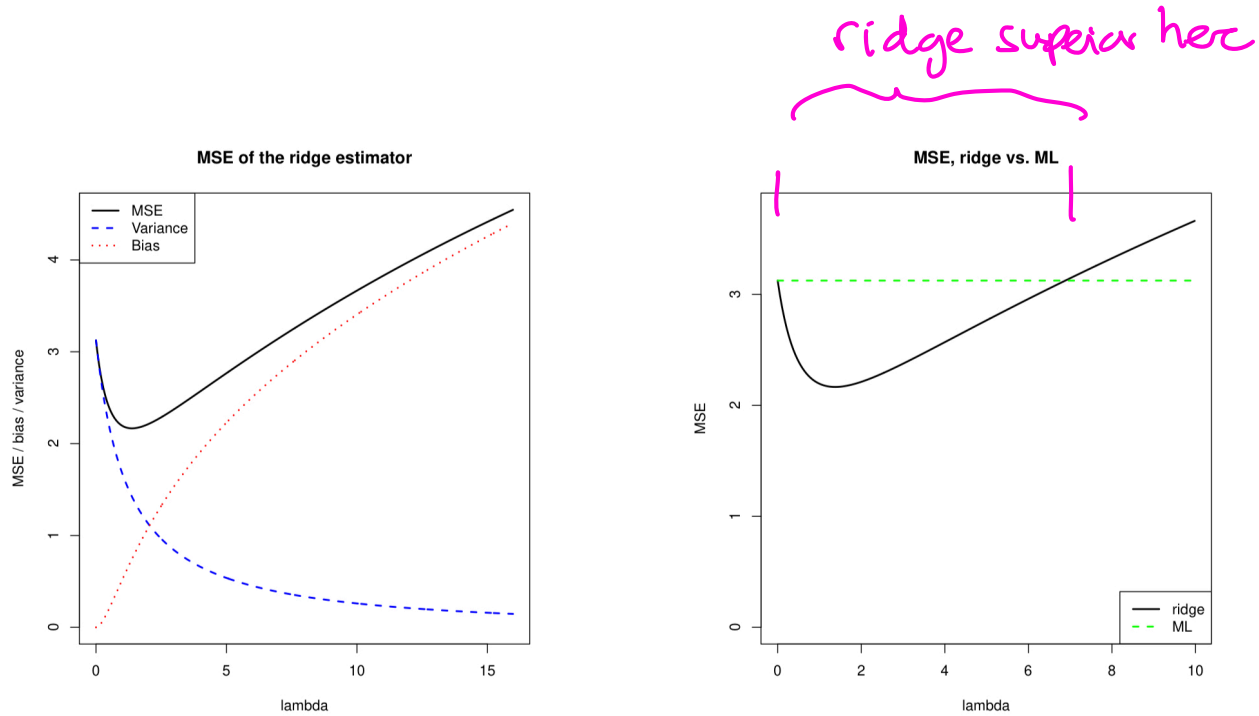


Figure 1.3: Left panel: mean squared error, and its 'bias' and 'variance' parts, of the ridge regression estimator (for artificial data). Right panel: mean squared error of the ridge and ML estimator of the regression coefficient vector (for the same artificial data).

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

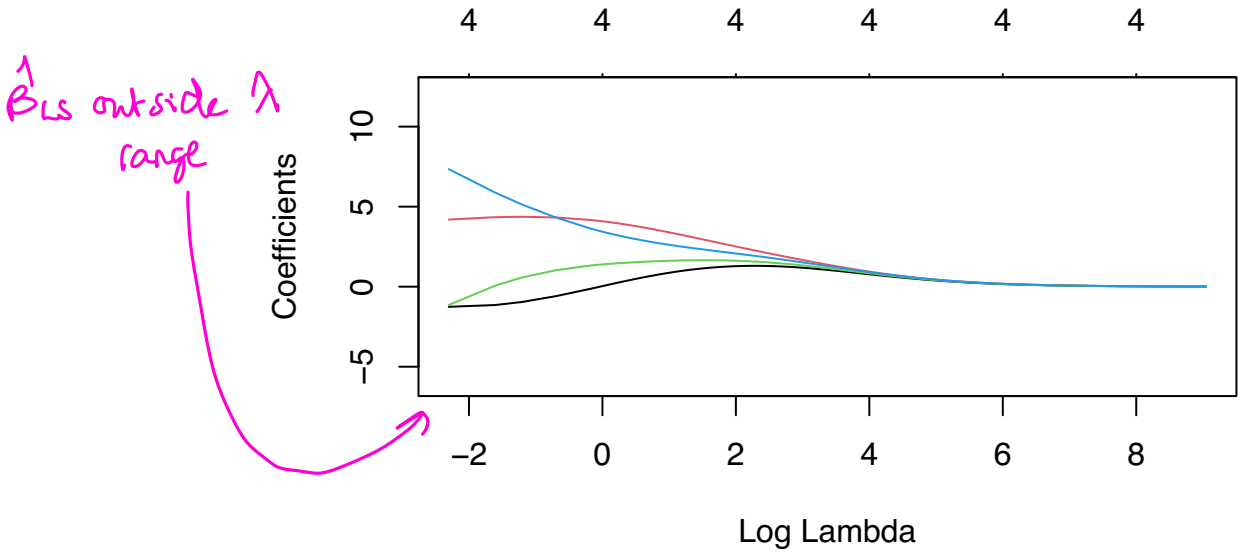
Model selection

To choose the optimal penalty parameter λ cross-validation is the default method in use. ELS recommends to either

- choose the λ corresponding to the smallest CV error
- or first find the λ with the smallest CV-error, and then record the estimated standard error of the CV-error at this value, and then choose the largest λ 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.

```
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

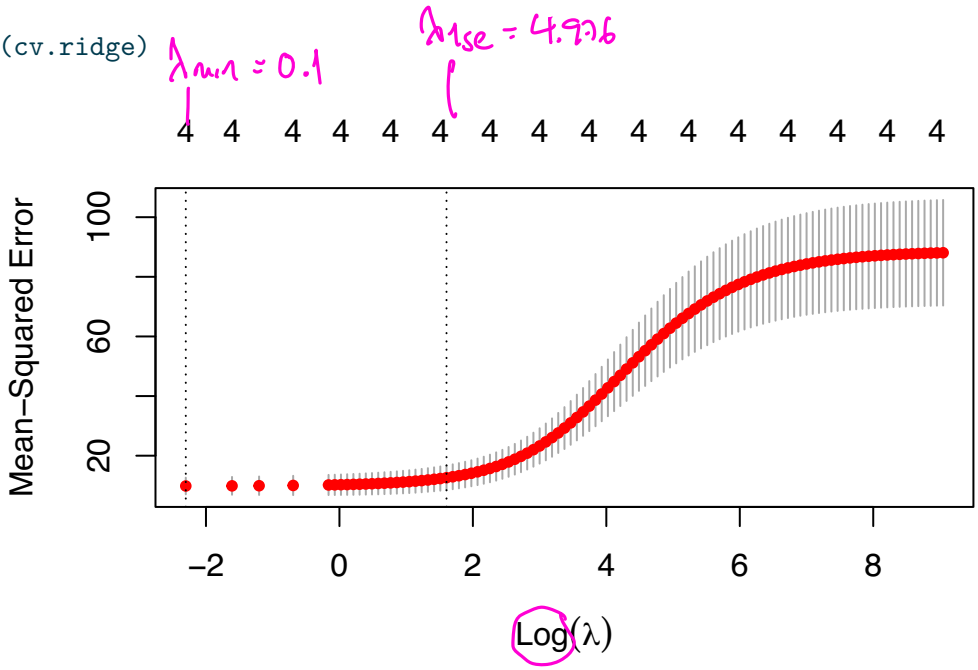
Explain what you see!

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

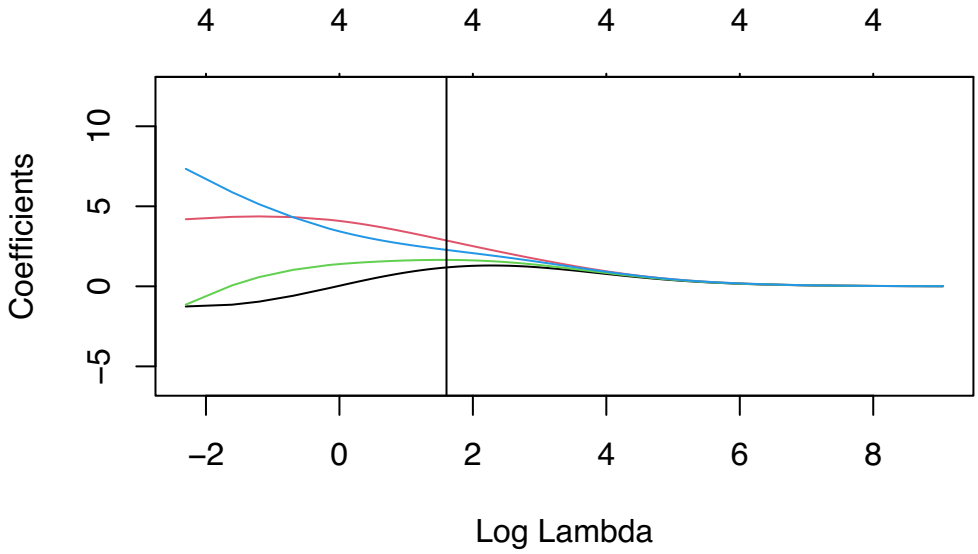
```
Call:  cv.glmnet(x = x, y = y, lambda = newlambda, alpha = 0)

Measure: Mean-Squared Error

    Lambda Index Measure    SE Nonzero
min  0.100   104   9.835 2.790       4
1se  4.976    81  12.616 4.124       4
```

← since 10-fold CV may change if rerun (if seed changed)

```
#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)
```




```
# 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) -2.097194e-15
TankTemp     1.181294e+00
GasTemp      2.863238e+00
TankPres     1.651347e+00
GasPres      2.276497e+00
```

compare with full model
→ observe shrinkage!

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

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

```
full$coeff
```

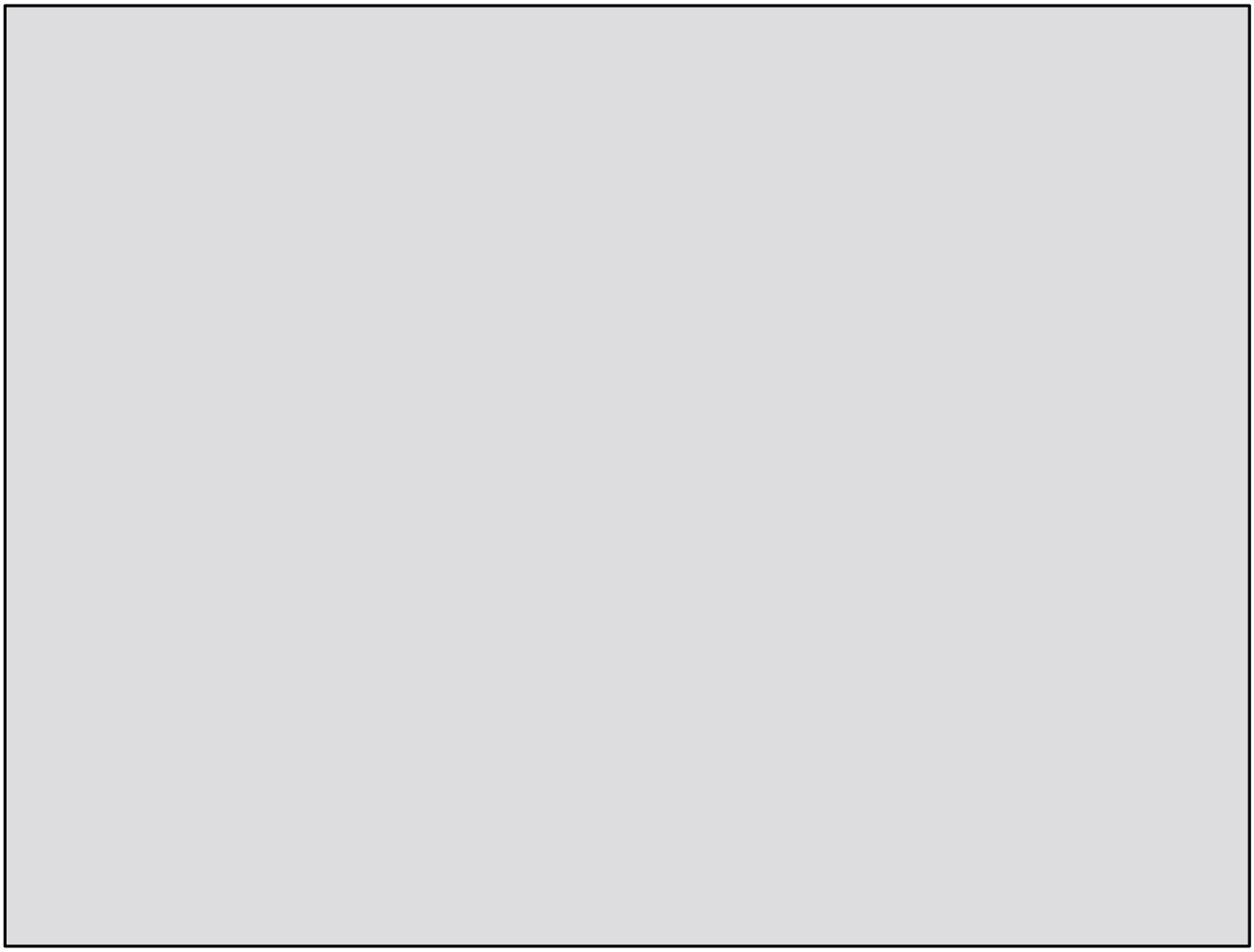
```
  (Intercept)      TankTemp       GasTemp      TankPres       GasPres
3.232869e-16 -5.581796e-01  3.395309e+00 -6.273748e+00  1.249044e+01
```

```
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
```



INSIGHTS FROM SVD

ELS 3.41

$(N-1)S = X^T X$ for centered data

$X = U D V^T$

columns are eigenvectors of XX^T (colspace of X)

D: sqrt of eigenvalue of $X^T X$ (or XX^T since equal)

ordered such that $d_1 \geq d_2 \geq \ldots \geq d_p \geq 0$

$D = \text{diag}(d_j)$

columns are eigenvectors of $X^T X$ (rowspace of X)

$X^T X = V D U^T U D V^T = V D^2 V^T$

$X\hat{\beta}_{LS} = X(X^TX)^{-1}X^TY = \dots$

$X\hat{\beta}(\lambda) = UDV^T(VD^2V^T + \lambda VV^T)^{-1}VDU^TY = \dots$

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

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

Ex: $\lambda = 1$

$d_j = 1 \Rightarrow \frac{d_j^2}{d_j^2+\lambda} = \frac{1}{2}$ ← shrink a lot

$d_j = 2 \Rightarrow = \frac{4}{5}$ ← shrink less

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 $\hat{\mathbf{y}} = \mathbf{S}y$ as

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

For ridge regression our linear smoother is

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

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

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

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

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.

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?)

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 the variance of $\tilde{\beta}$ exceeds *in a positive definite ordering sense* that of $\hat{\beta}$, and written $\mathrm{Var}(\tilde{\beta}) \succeq \mathrm{Var}(\hat{\beta})$.

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 4, 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 $\mathbf{Y}$

Very useful exercise!

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 λ gives the minimum value of the MSE for the ridge regression estimator?