Chapter 1: Variable Selection and Model Selection

Maarten Jansen

PART 1: MULTIPLE LINEAR REGRESSION

The general linear model (1)

\[ Y_i = \beta_0 + \sum_{j=1}^{p-1} \beta_j x_{i,j} + \epsilon_i \]

With:

- \( Y_i \) the \( i \)th observation of the response variable
  - dependent variable
  - explained, observed variable
- \( x_{i,j} \) the (observed) \( i \)th value of the \( j \)th covariate also called
  - independent variable
  - regression variable
  - explanatory variable
  - predictor, predicting variable
  - control variable

Overview

1. Multiple linear regression (Slide 2)
2. Variable and model selection (Slide 60)
3. Model averaging (Slide 127)
4. High dimensional model selection (Slide 108)
Examples, specific cases

1. **Polynomial regression**

   \[ Y_i = \beta_0 + \sum_{j=1}^{p-1} \beta_j x_i^j + \epsilon_i \]

   So \( x_{i,j} = x_i^j \).

   There is only one explanatory variable in strict sense (\( x \) with observations \( x_i \)), but the response depends on several powers of \( x \).

2. **Models with interaction**

   Suppose: 2 explanatory variables, \( x_1 \) and \( x_2 \), but \( Y \) also depends on the interaction between them.

   \[ Y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_{1,2} x_{i,1} x_{i,2} + \epsilon_i \]

   We then set \( x_3 = x_1 x_2 \).

3. **One-way variance analysis (ANOVA)**

   \[ Y_{ij} = \mu_i + \epsilon_{ij} \]

   with \( i = 1, \ldots, k \) and \( j = 1, \ldots, n_i \).

   Note that in this model the observations have double indices.

   All observations at the same level \( \mu_i \) are only different by the observational noise, so in a regression model they should share common variables of the explanatory variables.

---

**Algebraic interpretation/Euclidean space**

Model

\[ Y = X\beta + \epsilon \]

We search for \( \hat{\beta} \) so that \( X\hat{\beta} \) is as close as possible to \( Y \) in quadratic norm.

**Algebraically** this means that we look for the minimum least squares solution to the overdetermined set of equations \( Y = X\beta \) has to be minimized.

Formally, find \( \hat{\beta} \) that minimizes \( \| Y - X\beta \|^2 \)

In a **Euclidean space** (i.e., a space where the notion of orthogonality can be defined using inner products) that means that the **residual vector** \( e = Y - X\hat{\beta} \) has to be orthogonal (perpendicular) to the approximation (or projection) \( X\hat{\beta} \) in the subspace spanned by \( X \).

On other words: The residual vector has minimum norm \( \Leftrightarrow \) it is orthogonal to the estimator vector.

---

**The normal equation and least squares (LS) estimator**

Developing the algebraic interpretation:

\[ (X\hat{\beta})^T (Y - X\hat{\beta}) = 0 \]

This is the case if \( X^T(Y - X\hat{\beta}) = 0 \)

from which the normal equation follows

\[ \hat{\beta} = (X^TX)^{-1}X^TY \]

Let \( \hat{\mu} \) be the estimator for \( \mu = X\beta \), then

\[ \hat{\mu} = X(X^TX)^{-1}X^TY \]
Special case: \( p = 1 \), simple linear regression

\[
Y_i = \beta_0 + \beta_1 x_{i,1} + \varepsilon_i, \quad i = 1, \ldots, n
\]

Simple linear regression: \( x_0 = 1 \) and \( x_1 = x \).

Explicit formulas for LS estimators

\[
\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (Y_i - \bar{Y})(x_{i,1} - \bar{x}_1)}{\sum_{i=1}^{n} (x_{i,1} - \bar{x}_1)^2}
\]

\[
\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}_1
\]

Least squares as an orthogonal projection

The prediction \( \hat{\mu} \) is an orthogonal Projection with projection matrix

\[
P = X(X^T X)^{-1}X^T
\]

so that

\[
\hat{\mu} = PY
\]

Orthogonality \( P^T = P \)

Projection \( PP = P \) (idempotence)

\[
P\mu = \mu \quad PX = X
\]

Interpretation: columns of \( X \) contain eigenvectors with eigenvalue 1. If \( X^T Y = 0 \), then \( PY = 0 \) (from definition of \( P \)). These are eigenvectors with eigenvalue 0.

Next slides discuss projections and orthogonal projections in general

Projections

\( P \) is a projection matrix \( \Leftrightarrow PP = P \) i.e., \( P \) is idempotent

Properties

1. \( P \) has eigenvalues 0 and/or 1
   
   Indeed, let \( PX = \lambda x \) with \( x \) non-trivial, then \( \lambda^2 x = PPx = PX = \lambda x \), hence \( \lambda^2 = \lambda \)

2. There exists an orthogonal \( Q \) so that \( P = Q \begin{bmatrix} I & A \\ 0 & 0 \end{bmatrix} Q^T \)

\[
P = E \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}^{-1} = QR \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} R^{-1} Q^T = Q \begin{bmatrix} I & R_{12} (R^{-1})_{12} \\ 0 & 0 \end{bmatrix} Q^T
\]

using the notations of submatrices \( R = \begin{bmatrix} R_{11} & R_{12} \\ 0 & R_{22} \end{bmatrix} \) and similar for \( R^{-1} \)

3. Singular values of \( P \) are either 0 or larger than 1

   Indeed, singular values of \( PP^T = Q \begin{bmatrix} I + AA^T & 0 \\ 0 & 0 \end{bmatrix} \) which are either 0 or 1+

Orthogonal projections

Projection \( P \) is an orthogonal projection \( \Leftrightarrow (I - P)^T P = 0 \)

So, the matrix \( P \) itself is NOT orthogonal: \( P^T P \neq I \)

Projection \( P \) is an orthogonal projection \( \Leftrightarrow P^T = P \)

Indeed, \( (I - P)^T P = 0 \Leftrightarrow P^T P = P \)

But then \( P^T = (P^T P)^T = P^T P = P \)

Hence \( P \) must be symmetric
Residuals

The vector of **Residuals** is $\hat{e} = Y - \hat{\mu} = (I - P)Y$ (See also slide 6)

The normal equation finds the estimator that minimizes the mean squared residual $e^T e / n = \|e\|^2 / n = \frac{1}{n} \sum_{i=1}^{n} e_i^2$

- The matrix $I - P$ is idempotent: $(I - P)(I - P) = I - P$
- This matrix has rank $n - p$: all linear combinations of the columns of $X$ belong to the kernel of the matrix, since $(I - P)X = 0$
- An idempotent matrix with rank $k$ has $k$ independent eigen vectors with eigen value 1 and $n - k$ independent eigen vectors with eigen value 0. The latter are the columns of $X$
- The matrix $I - P$ is symmetric, so eigen vectors corresponding to eigen value 1 are orthogonal to eigen vectors with eigen value 0. Within each of the two groups of eigen vectors, orthogonalisation (using Gram-Schmidt) is straightforward.
- So: there exists an orthogonal transform $U$ so that $I - P = U^T \Lambda U$ with $\Lambda = \text{diag}(1, \ldots, 1, 0, \ldots, 0)$

Gram-Schmidt-orthogonalisation

Let $x_j, j = 0, \ldots, p - 1$ be the columns of $X$, then define

\[
q_0 = x_0 = 1 \\
q_1 = x_1 - \frac{\langle x_1, q_0 \rangle}{\|q_0\|^2} \cdot q_0 \\
\ldots \\
q_j = x_j - \sum_{k=0}^{j-1} \frac{\langle x_j, q_k \rangle}{\|q_k\|^2} \cdot q_k
\]

This is: $q_j$ is the residual of orthogonal projection of $x_j$ onto $q_0, \ldots, q_{j-1}$

Looks like linear regression within the covariate matrix $X$

The new covariates $q_j$ are orthogonal, but not orthonormal (they can be easily normalised)

Orthogonalisation

If $X$ has **orthogonal columns**, then $X^T X = I_p$, and thus the projection becomes

\[
\hat{\mu} = X \cdot X^T \cdot Y
\]

The $i$th component is then

\[
\hat{\mu}_i = \sum_{j=0}^{n-1} \sum_{k=1}^{p} \langle x_{i,j}, q_{k,j} \rangle \cdot Y_k
\]

The computation of all $n$ components of $\hat{\mu}$ can proceed in parallel, without numerical errors progressing from one component to another.

Otherwise, we **Orthogonalise** $X$

Gram-Schmidt in matrix form

We have

\[
x_j = q_j + \sum_{k=0}^{j-1} \frac{\langle x_j, q_k \rangle}{\|q_k\|^2} \cdot q_k
\]

Apply scalar product with $q_j$ on both sides to find that $\langle x_j, q_j \rangle / \|q_j\|^2 = 1$.

Hence

\[
x_j = \sum_{k=0}^{j-1} \frac{\langle x_j, q_k \rangle}{\|q_k\|^2} \cdot q_k
\]

In matrix-form: **QR-decomposition** $X = QR$ with the columns $Q$ equal to $q_j$, and $R$ an upper triangular matrix, with entries

\[
R_{k,j} = \frac{\langle x_j, q_k \rangle}{\|q_k\|^2}
\]

$R \in \mathbb{R}^{p \times p}$ and $R$ is invertible.

**Interpretation** Every column of $X$ (every covariate $x_j$) can be written as a linear combination of orthogonal covariates, which are the columns of $Q$. (Right-multiplication with $R$)
Regression after orthogonalisation

Plugging in into the solution of the normal equation:
\[
\hat{\mu} = QR(R^T Q R)^{-1} R^T Q \cdot Y = Q(Q^T Q)^{-1} Q^T \cdot Y
\]
This is a projection onto the orthogonal (but not orthonormal) basis \(Q\).

Written out:
\[
\hat{\mu} = \sum_{k=0}^{p-1} \frac{(Y, q_k)}{\|q_k\|^2} \cdot q_k
\]

The vector of coefficients \(\hat{\gamma}\) with
\[
\hat{\gamma}_k = \frac{(Y, q_k)}{\|q_k\|^2}
\]
satisfies \(\hat{\mu} = Q \hat{\gamma}\) and \(\hat{\gamma} = (Q^T Q)^{-1} Q^T \cdot Y\)

Relation between \(\hat{\gamma}\) and \(\hat{\beta}\):
\[
\hat{\beta} = (X^T X)^{-1} X^T \cdot Y = (R^T Q^T Q R)^{-1} R^T Q^T \cdot Y = R^{-1} (Q^T Q)^{-1} Q^T \cdot Y
\]
\[
\hat{\beta} = R^{-1} \hat{\gamma} \quad \hat{\gamma} = R \hat{\beta} \quad \text{(Left-multiplication with } R\text{)}
\]

Tools for statistical exploration

In order to further explore the properties of the least squares estimator, we need to establish some classical results in multivariate statistics

1. Covariance matrix
2. Multivariate normal distribution
3. Trace of a matrix

Orthogonalisation details for \(p = 1\), simple linear regression

Model
\[
Y_i = \beta_0 + \beta_1 x_{i,1} + \varepsilon_i, \quad i = 1, \ldots, n
\]

Simple linear regression: \(x_0 = 1\) and \(x_1 = x\).

\[
q_1 = x_1 - \frac{(x_1, q_0)}{\|q_0\|^2} \cdot q_0 = x_1 - \frac{(x_1, 1)}{1^2} \cdot 1 = x_1 - \frac{\sum_{i=1}^n x_{i,1}}{n} \cdot 1 = x_1 - \bar{x}_1 \cdot 1
\]

\[
R = \begin{bmatrix}
\frac{1}{n} & \bar{x}_1 \\
0 & 1
\end{bmatrix} \quad R^{-1} = \begin{bmatrix}
\frac{n}{1} & -n\bar{x}_1 \\
0 & 1
\end{bmatrix}
\]

\[
\hat{\gamma}_0 = \frac{1}{n} \sum_{i=1}^n x_{i,0} \quad \hat{\gamma}_1 = \frac{(q_0, Y)}{\|q_0\|^2} = \frac{(x_1 - \bar{x}_1 \cdot 1, Y)}{\|x_1 - \bar{x}_1 \cdot 1\|^2}
\]

Since the second row of \(R^{-1}\) equals \(\begin{bmatrix} 0 & 1 \end{bmatrix}\), we find immediately that
\[
\hat{\beta}_1 = \frac{(x_1 - \bar{x}_1, 1, Y)}{\|x_1 - \bar{x}_1 \cdot 1\|^2}
\]

which corresponds to the classical result, page 8

Tool 1: Covariance matrix

Definition

Let \(X\) be a \(p\)-variate random vector with joint density function \(f_X(x)\), joint (cumulative) distribution function \(F_X(x)\), and joint characteristic function
\[
\phi_X(t) = E(e^{it^T X})
\]

\[
\mu = E(X)
\]

\[
\Sigma_X = E[(X - \mu)(X - \mu)^T] \quad \text{We have} \quad (\Sigma_X)_{ij} = \text{cov}(X_i, X_j)
\]

\(\Sigma_X\) is symmetric
If $Y = AX$ Then $\Sigma_Y = A \Sigma_X A^T$

Proof:
It holds that $\mu_Y = A \mu_X$, so

$$\Sigma_Y = E [(Y - \mu_Y)(Y - \mu_Y)^T] = E [(A(X - \mu_X))(A(X - \mu_X))^T]$$

So, with $A_i$ the $i$th column of $A$, we have

$$\text{cov}(Y_i, Y_j) = A_i^T \Sigma_X A_j = A_T \Sigma_X A_i$$

Special case: matrix $A$ has just one row: $A = a^T$, then $Y = a^T X$ and $\sigma_Y^2 = a^T \Sigma_X a$

A symmetric matrix $S$ is called **positive (semi-)definite** if $x^T S x > (\geq) 0$ for every vector $x \neq 0$

A positive-definite matrix always has real, positive eigenvalues

$\Sigma_X$ is positive semi-definite, since if $a^T \Sigma_X a < 0$, then $Y = a^T X$ would have a negative variance

We further assume that $\Sigma_X$ is positive definite.

**Tool 2: The multivariate normal distribution**

**Definition: multivariate normal distribution**

$X \sim N(\mu, \Sigma_X) \iff Z \sim N(0, 1)$

As the components of $Z$ are uncorrelated, and (here) normally distributed, they must me mutually independent.

The joint density function

$$f_X(x) = \frac{1}{(2\pi)^{p/2} \sqrt{\det(\Sigma_X)}} e^{-\frac{1}{2}(x - \mu)^T \Sigma_X^{-1}(x - \mu)}$$

The characteristic function

$$\phi_X(t) = \exp \left( it^T \mu - \frac{1}{2} t^T \Sigma_X t \right)$$
### Properties of the multivariate normal distribution

**Property 1**
If $X \sim N(\mu, \Sigma_X)$, then $(X - \mu)^T \Sigma_X^{-1} (X - \mu) \sim \chi^2_p$

This follows from

$$(X - \mu)^T \Sigma_X^{-1} (X - \mu) = (X - \mu)^T \Sigma^{-1} X = Z^T Z = \sum_{i=1}^p Z_i^2$$

and $Z_i \sim \chi^2_1$ and $Z_i$ mutually independent

**Property 2**

If $X \sim N(\mu, \Sigma_X)$, then for $C \in \mathbb{R}^{q \times p}$ with $q \leq p$: $CX \sim N(C\mu, C\Sigma_X C^T)$

**Property 3**

$$X \sim N(\mu, \Sigma_X) \iff a^T X \sim N(a^T \mu, a^T \Sigma_X a)$$

### Marginal and conditional covariance

(further details of previous slide)

Suppose $X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$ with mean $\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$ and covariance matrix $\Sigma_X = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$ and $\det(\Sigma_{22}) \neq 0$, then

**Marginal covariance**

$$\text{cov}(X_1) = \Sigma_{11}$$

**Conditional covariance**

$$\text{cov}(X_1|X_2) = \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21}$$

This expression is known as the **Schur-complement**.

It holds that $\text{cov}(X_1|X_2) = \Sigma_{12} \Sigma_{22}^{-1} = [\Sigma_{11}^{-1}]_{11}$

### Proof

Construct $L = \begin{bmatrix} l_{11} & 0 \\ -\Sigma_{22}^{-1} \Sigma_{21} & l_{22} \end{bmatrix}$, then $L^{-1} = \begin{bmatrix} l_{11} & 0 \\ +\Sigma_{22}^{-1} \Sigma_{21} & l_{22} \end{bmatrix}$

and

$$\Sigma_X L = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \begin{bmatrix} l_{11} & 0 \\ -\Sigma_{22}^{-1} \Sigma_{21} & l_{22} \end{bmatrix} = \begin{bmatrix} \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

$$= \begin{bmatrix} l_{11} & 0 \\ 0 & l_{22} \end{bmatrix}$$

So, $\Sigma_X = \begin{bmatrix} l_{11} & \Sigma_{12} \Sigma_{22}^{-1} \\ 0 & l_{22} \end{bmatrix} \begin{bmatrix} \Sigma_{12} & 0 \\ 0 & \Sigma_{22} \end{bmatrix} \begin{bmatrix} l_{11} & 0 \\ 0 & l_{22} \end{bmatrix} = \begin{bmatrix} l_{11} & 0 \\ -\Sigma_{21} \Sigma_{22}^{-1} \Sigma_{12} & l_{22} \end{bmatrix}$

and

$$\Sigma_X^{-1} = \begin{bmatrix} l_{11} & \Sigma_{12} \Sigma_{22}^{-1} \\ 0 & l_{22} \end{bmatrix} \begin{bmatrix} l_{11} & 0 \\ 0 & l_{22} \end{bmatrix} = \begin{bmatrix} l_{11} & 0 \\ 0 & l_{22} \end{bmatrix}$$

$$= \begin{bmatrix} l_{11} & \Sigma_{12} \Sigma_{22}^{-1} \\ 0 & l_{22} \end{bmatrix} \begin{bmatrix} l_{11} & 0 \\ 0 & l_{22} \end{bmatrix} = \begin{bmatrix} l_{11} & 0 \\ 0 & l_{22} \end{bmatrix}$$

Note also that $\Sigma_{21} = \Sigma_{12} \Sigma_{22}^{-1}$
### Tool 3: Trace of a matrix

\[
\text{Tr}(A) = \sum_{i=1}^{p} a_{ii}
\]

\[
\text{Tr}(AB) = \text{Tr}(BA) \quad (A \in \mathbb{R}^{p \times q}, B \in \mathbb{R}^{q \times p})
\]

Holds for both square \((p = q)\) and rectangular matrices with matching sizes.

**Proof:**
\[
\text{Tr}(AB) = \sum_{i=1}^{p} \sum_{j=1}^{q} A_{ij}B_{ji} = \sum_{j=1}^{q} \sum_{i=1}^{p} B_{ji}A_{ij} = \sum_{j=1}^{q} \sum_{i=1}^{p} (BA)_{ij} = \text{Tr}(BA)
\]

**Corollary A. B. C square matrices, then:** \(\text{Tr}(ABC) = \text{Tr}(BCA) = \text{Tr}(CAB)\)

Trace remains unchanged under cyclic permutation, so NOT under every permutation:
\(\text{Tr}(ABC) \neq \text{Tr}(BAC)\)

Holds also for more than 3 matrices and also for rectangular matrices with matching dimensions.

\[
\text{Tr}(A + B) = \text{Tr}(A) + \text{Tr}(B)
\]

Attention: \(\text{Tr}(A \cdot B) \neq \text{Tr}(A) \cdot \text{Tr}(B)\) but \(\det(A \cdot B) = \det(A) \cdot \det(B)\)

**Trace and eigenvalues**

If \(A = E \Lambda E^{-1}\), with \(\Lambda = \text{diag}(\lambda_i)\), then
\[
\text{Tr}(A) = \sum_{i=1}^{n} \lambda_i \quad \det(A) = \prod_{i=1}^{n} \lambda_i
\]

---

### Maximum likelihood for regression

Model \(Y = X\beta + \epsilon\) (random variable \(Y\) instead of \(X\) in previous slides)

Assume \(\epsilon \sim \text{N.I.D.}(0, \sigma^2)\), then joint observational density is multivariate normal, more precisely

\[
L(\beta, \sigma^2; Y) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2} (Y - X\beta)^T(Y - X\beta)\right)
\]

To be maximised over \(\beta\) and \(\sigma^2\).

\[
\log L(\beta, \sigma^2; Y) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(Y - X\beta)^T(Y - X\beta)
\]

Dependence on \(\beta\) only through second term, so \(\beta\) maximises \(L(\beta, \sigma^2; Y)\) or

\[
\log L(\beta, \sigma^2; Y)\] iff \(\beta\) minimises

\[
(Y - X\beta)^T(Y - X\beta) = \|Y - X\beta\|_2^2
\]

This is least squares.

Conclusion: for normal errors is the least squares solution the maximum likelihood solution.

---

### Property 4: best linear unbiased estimator (BLUE)

**Suppose**

- **Zero mean errors** \(E(\epsilon) = 0\)
- **Uncorrelated, homoscedastic errors** \(\Sigma_\epsilon = \sigma^2 I\) i.e.: \(\text{var}(\epsilon_i) = \sigma^2\) and \(\text{cov}(\epsilon_i, \epsilon_j) = 0\) (not necessarily independent)
- Normality is not necessary

While \(\tilde{\beta} = (X^T X)^{-1}X^TY\), let \(\tilde{\beta} = AY = [(X^T X)^{-1}X^T + \Delta] Y\), be another unbiased, linear estimator, then:

#### Gauss-Markov theorem

For any choice of \(c^T\), under the assumptions above, \(\text{MSE}(c^T \tilde{\beta}) \geq \text{MSE}(c^T \beta)\)

In other words, for any \(A\) so that \(E(AY) = \beta\) (unbiased), it holds that

\[
\Sigma_\beta - \Sigma_{\tilde{\beta}} = \text{positive semi-definite}
\]

**Note:**

\[
\text{MSE}(\beta) = E \left[ (\tilde{\beta} - \beta)^2 \right] \quad \text{MSE}(\tilde{\beta}) = \left[ E(\tilde{\beta} - \beta)^2 \right] + \text{var}(\tilde{\beta})
\]
**Gauss-Markov theorem: proof**

As \(E(AY) = \beta\), we have to prove that \(\text{var}(c^T \hat{\beta}) \geq \text{var}(c^T \hat{\beta})\)

Using the linearity, we find, for any value of \(\beta\)

\[
\beta = E(AY) = AE(Y) = A(X\beta),
\]

hence \(AX = I\) and so \(\Delta X = O\) for \(\Delta = A - (X^TX)^{-1}X^T\)

The covariance matrix of \(\hat{\beta}\) is \(\sigma^2(X^TX)^{-1}\) (slide 30), so

\[
\text{var}(c^T \hat{\beta}) = \sigma^2 c^T (A^T - (X^TX)^{-1})c
\]

We can write

\[
A^T - (X^TX)^{-1} = [(X^TX)^{-1}X^T + \Delta] [(X^TX)^{-1}X^T + \Delta]^T - (X^TX)^{-1}
\]

\[
= (X^TX)^{-1}X^T \Delta^T + \Delta X(X^TX) + \Delta \Delta^T
\]

The quadratic form \(\Delta \Delta^T\) is always positive, semi-definite, thereby completing the proof

---

**Remark: the expected squared loss**

**Definition: Risk or expected squared loss**

\[
R(\hat{\beta}) = \frac{1}{n} E (||\hat{\beta} - \beta||^2)
\]

This definition satisfies \(R(\hat{\beta}) = \frac{1}{n} \text{tr}(\Sigma_{\hat{\beta}}) + \frac{1}{n} ||E\hat{\beta} - \beta||^2\)

We know: \(\Sigma_{\hat{\beta}} - \Sigma_{\tilde{\beta}}\) is positive semi-definite, thereby completing the proof

This implies as a necessary condition:

\[
\Sigma_{\hat{\beta}} - \Sigma_{\tilde{\beta}}\text{ pos.def} \Rightarrow \text{tr} (\Sigma_{\hat{\beta}} - \Sigma_{\tilde{\beta}}) \geq 0.
\]

Hence, for unbiased estimators:

\[
\Sigma_{\hat{\beta}} - \Sigma_{\tilde{\beta}}\text{ pos.def} \Rightarrow R(\hat{\beta}) \leq R(\tilde{\beta})
\]

Risk is sometimes referred to as MSE, but working with MSE of any linear combination \(MSE(c^T \hat{\beta})\) as on slide 31 puts a stronger condition

On the other hand, if we can find an example where \(R(\tilde{\beta}) \geq R(\hat{\beta})\), then \(\Sigma_{\hat{\beta}} - \Sigma_{\tilde{\beta}}\) cannot possibly be positive definite

---

**Stein’s phenomenon**

The proof of the Gauss-Markov theorem hinges on the fact that

\[(X^TX)^{-1}X^T \Delta^T = 0\]. This is due to the linearity and unbiasedness of the estimator \(AY\) and to the specific form of the LS estimator

\(\hat{\beta} = (X^TX)^{-1}X^TY\).

Stein’s phenomenon

- Biased estimators may have uniformly (i.e., for any true value) lower risk

Stein’s example

- \(Y = \mu + \varepsilon\), we want to estimate \(\mu\) (i.e., no regression, just a signal-plus-noise model)
- \(\varepsilon \sim N_n(\mu, \sigma^2 I)\): noise is normal, homoscedastic and uncorrelated (hence i.i.d. normal)

The **BLUE estimator** is \(\hat{\mu} = Y\).

The risk (slide 33) is \(R(Y) = \sigma^2\)

A biased, shrinkage estimator

\[
\hat{\mu}_i = Y_i(1 - \alpha/||Y||^2)
\]

---

Now we use **Stein’s Lemma**

If \(X \sim N(\mu, \sigma^2 I)\) then \(E[(X - \mu)h(X)] = E[h'(X)]\) with \(h(x)\) continuous

Here: \(X = Y_i, h(X) = h_i(Y_i) = Y_i/||Y||^2\), so \(h'(X) = \frac{\partial h_i(Y_i)}{\partial Y_i} = ||Y||^2 / ||Y||^4\)

As a vector with \(h(Y) = Y / ||Y||^2\)

\[
E[(Y - \mu)^T h(Y)] = \sum_{i=1}^n E \left[ \frac{\partial h_i(Y_i)}{\partial Y_i} \right] = \sum_{i=1}^n E [h_i(Y_i)]
\]

\[
= \sum_{i=1}^n E \left[ \frac{1}{||Y||^2} - \frac{2Y_i}{||Y||^4} \right] = E \left[ \frac{n}{||Y||^2} - \frac{2}{||Y||^2} \right]
\]

Suppose \(n > 2\) and take \(\alpha = n - 2\), then \(R(\hat{\mu}) \leq \sigma^2 = R(Y)\)
Example 2: ridge regression/Tikhonov regularisation

If \( X \) has nearly linearly dependent columns, the \( X^T X \) is nearly singular

Variance of \( \hat{\beta} \) can be arbitrarily large (see p. 30)

We give up unbiasedness for the sake of variance reduction: regularisation of a (nearly) singular system

The least squares problem \( \min_\beta \| Y - X\beta \|^2 \) leads to the singular normal equation \( (X^T X)\hat{\beta} = X^T Y \)

We regularise \( \min_\beta \| Y - X\beta \|^2 + \| L \beta \|^2 \)

Very often, one takes \( L = \sqrt{\Lambda} I \), so \( \min_\beta \| Y - X\beta \|^2 + \lambda \| \beta \|^2 \)

The solution is

\[ \hat{\beta} = (X^T X + L^T L)^{-1} X^T Y \] or \[ \hat{\beta} = (X^T X + \lambda I)^{-1} X^T Y \]

Covariance matrix of the regularised solution

Consider the regularized estimator as a function of \( \lambda \):

\[ \hat{\beta}(\lambda) = (X^T X + \lambda I)^{-1} X^T Y \]

Then \( \Sigma_{\hat{\beta}(\lambda)} = \sigma^2(X^T X + \lambda I)^{-1}(X^T X)(X^T X + \lambda I)^{-1} \)

Let \( X^T X = \Lambda U \Lambda^T \), where \( \Lambda \) is a diagonal matrix, then

\[ \Sigma_{\hat{\beta}(\lambda)} = \sigma^2 U (\Lambda + \lambda I)^{-1} U^T \Lambda U^T (\Lambda + \lambda I)^{-1} U \sigma^2 U(\Lambda + \lambda I)^{-2} \Lambda U^T \]

The effect of the regularization can be described by

\[ \Sigma_{\hat{\beta}(0)} - \Sigma_{\hat{\beta}(\lambda)} = \sigma^2 U [\Lambda^{-2} - (\Lambda + \lambda I)^{-2}] \Lambda U^T \]

The diagonal matrix \( D = [\Lambda^{-2} - (\Lambda + \lambda I)^{-2}] \Lambda \) has elements

\[ D_{ii} = \frac{1}{\Lambda_{ii}} - \frac{\Lambda_{ii}}{(\Lambda_{ii} + \lambda)^2} \geq 0 \]

Hence, \( \Sigma_{\hat{\beta}(0)} - \Sigma_{\hat{\beta}(\lambda)} \) is positive definite, meaning that, for any \( c \)

\[ \text{var}(c^T \hat{\beta}(0)) = c^T \Sigma_{\hat{\beta}(0)} c \geq c^T \Sigma_{\hat{\beta}(\lambda)} c = \text{var}(c^T \hat{\beta}(\lambda)) \]

The bias of the regularised solution

We check if/when the decrease in variance dominates the increase in bias

\[ \text{bias}(\hat{\beta}(\lambda)) = E(\hat{\beta}(\lambda)) - \beta = (X^T X + \lambda I)^{-1} X^T X \beta - \beta \]

\[ = (X^T X + \lambda I)^{-1} [X^T X - (X^T X + \lambda I)] \beta \]

\[ = -\lambda (X^T X + \lambda I)^{-1} \beta \]

\[ \text{bias}(c^T \hat{\beta}(\lambda)) = -\lambda c^T (X^T X + \lambda I)^{-1} \beta \]

depends on \( c \) and \( \beta \)

No general conclusion

We investigate the risk of the regularised estimator. If the risk for a nonzero \( \lambda \) is smaller than for \( \lambda = 0 \), then \( \lambda = 0 \) cannot be possibly minimise

\( \text{MSE}(c^T \hat{\beta}(\lambda)) \) for any \( c \) (see p.33)

The risk of a ridge regression

\[ R(\hat{\beta}(\lambda)) = \frac{1}{n} \text{Tr} \left( \Sigma_{\hat{\beta}(\lambda)} \right) + \frac{1}{n} \| E \hat{\beta}(\lambda) - \beta \|^2 \]

\[ = \frac{\sigma^2}{n} \sum_{i=1}^{n} \frac{\Lambda_{ii}}{(\Lambda_{ii} + \lambda)^2} + \frac{1}{n} \| (X^T X + \lambda I)^{-1} \beta \|^2 \]

\[ \leq \frac{\sigma^2}{n} \sum_{i=1}^{n} \frac{\Lambda_{ii}}{(\Lambda_{ii} + \lambda)^2} + \frac{\lambda^2}{n} \| (X^T X + \lambda I)^{-1} \|^2 \| \beta \|^2 \]

\[ = \frac{\sigma^2}{n} \sum_{i=1}^{n} \frac{\Lambda_{ii}}{(\Lambda_{ii} + \lambda)^2} + \frac{\lambda^2}{n} \| \beta \|^2 \sum_{i=1}^{n} \frac{1}{(\Lambda_{ii} + \lambda)^2} \]

\[ = \frac{1}{n} \sum_{i=1}^{n} \frac{\Lambda_{ii} \sigma^2 + \lambda^2 \| \beta \|^2}{(\Lambda_{ii} + \lambda)^2} \]

All terms in this sum have a negative derivative for \( \lambda \) in \( \lambda = 0 \), hence

For any vector \( \beta \), there exists a parameter \( \lambda > 0 \), so that \( R(\hat{\beta}(\lambda)) < R(\hat{\beta}(0)) \)
Variance estimation

\[ S^2 = \frac{(Y - X\hat{\beta})^T (Y - X\hat{\beta})}{n-p} = \frac{\hat{e}^T \hat{e}}{n-p} \quad (\hat{e} = \text{residual, see slide 12}) \]

Note that the first factor is transpose (↔ expression of covariance matrix)

\[ ES^2 = \sigma^2 \]

**Proof** We know \( \tilde{\mu} = X\tilde{\beta} = PY \), so
\[ Y - X\tilde{\beta} = (I - P)Y = (I - P)(X\beta + \varepsilon) = (I - P)e \]
and so
\[ E \left[ (Y - X\tilde{\beta})^T (Y - X\tilde{\beta}) \right] = E \left[ e^T (I - P)^T (I - P)e \right] = E \left[ e^T (I - P)e \right] = E \left[ \text{Tr}(e^T (I - P)e) \right] = \text{Tr} \left[ (I - P)E(ee^T) \right] = \text{Tr}(I - P)\sigma^2 = (n-p)\sigma^2 \]

(Using that scalar = trace(scalar) and then cyclic permutation within trace, see slide 28 + properties of \( P \), see 9)

### Distribution of the variance estimator

**Theorem**

\[ (n-p)S^2/\sigma^2 \sim \chi^2_{n-p} \]

- There exists an orthogonal transform \( U \) so that \( I - P = U^T A U \) with \( A = \text{diag}(1, \ldots, 1, 0, \ldots, 0) \) (see slide 12)
- We know from the proof of \( ES^2 = \sigma^2 \) that
  \[ (n-p)S^2/\sigma^2 = \frac{(Y - X\tilde{\beta})^T (Y - X\tilde{\beta})}{\sigma^2} = \frac{\hat{e}^T (I - P)\hat{e}}{\sigma^2} = \frac{\hat{e}^T \hat{e}}{\sigma^2}U^T \Lambda U \hat{e} \]
- \( \frac{\hat{e}^T \hat{e}}{\sigma^2} \) is a vector of independent, standard normal random variables:
  \( \Sigma_{\hat{e}/\sigma} = I \). After application of the orthogonal transform \( Z = U\hat{e} \) we have
  \( \Sigma_Z = UU^T = I \), so \( Z \) is also a vector of independent, standard normal random variables, \( Z_i \), so
  \( (n-p)S^2/\sigma^2 = \sum_{i=1}^{n-p} Z_i^2 \sim \chi^2_{n-p} \)

### The sum of squared residuals is a biased estimator

**Interpretation**

The mean squared residual \( e^T e/n \) underestimates the expected squared error \( \sigma^2 = E(e^2) \). This is because the least squares method minimizes the mean squared residual, and so even the true model \( \beta \) has a larger mean squared residual than the estimated model.

The minimum mean squared residual is always smaller than the average squared error in the true model.

### Hypothesis testing - likelihood ratio

We want to test if \( H_0 : A\beta = c \) with \( A \in \mathbb{R}^{q \times p} \) where \( q \leq p \) and rank of \( A \) equal to its number of rows \( q \)

**Example** \( c = 0 \) and the rows of \( A = \begin{bmatrix} I & 0 \end{bmatrix} \). This leads to the hypothesis test if the first \( q \) \( \beta_j \) equal zero (other subsets of \( \beta_j \) are possible)

As \( \sigma \) is unknown, it is a component in the parameter vector \( \theta = (\beta, \sigma^2) \) to be estimated.

Let \( \Omega = \{ \theta \} \) be the vector space of all values that the parameter vector can take and \( \Omega_0 = \{ \theta \in \Omega | A\beta = c \} \) the subspace of values that satisfy the null hypothesis.

**Likelihood ratio**

\[ \Lambda = \frac{\max_{\theta_0 \in \Omega_0} L(\theta_0; Y)}{\max_{\theta \in \Omega} L(\theta; Y)} \]
**Likelihood ratio**

**Theorem** (from mathematical statistics)

If \( n \to \infty \), then 
\[
-2 \log \Lambda \xrightarrow{\text{d}} \chi^2_{\nu - \nu_0}
\]

\( \nu \) is the dimension of the space \( \Omega \), \( \nu_0 \) is the dimension of \( \Omega_0 \): in our case
\( \nu = p \), \( \nu_0 = p - q \)

**Maximum likelihood estimator in multiple regression**

\[
L(\beta; \sigma^2; Y) = \frac{1}{(2\pi \sigma^2)^n/2} e^{-1/2 (Y - X\beta)^T (Y - X\beta)}
\]

\[
\log L(\beta; \sigma^2; Y) = -\frac{n}{2} \log (2\pi \sigma^2) - \frac{1}{2\sigma^2} (Y - X\beta)^T (Y - X\beta)
\]

Taking derivatives for \( \sigma^2 \) yields
\[
\hat{\sigma}^2 = \frac{1}{n} (Y - X\hat{\beta})^T (Y - X\hat{\beta}) = \frac{n}{n - p} S^2 \]  
where \( \hat{\beta} \) follows from taking derivatives for \( \beta \) (least squares as discussed before; but we don’t need this for the subsequent analysis)

Plugging in leads to
\[
L(\hat{\beta}, \hat{\sigma}^2; Y) = \frac{1}{(2\pi \hat{\sigma}^2)^{n/2}} e^{-n/2}
\]

**Variance estimator under reduced model (1)**

Starting from the reduced model \( H_0 : A\beta = c \) with \( A \in \mathbb{R}^{q \times p} \) where \( q \leq p \) and rank of \( A \) equal to number of rows \( q \)

If \( H_0 \) holds, then
\[
\hat{\beta}_0 = \hat{\beta} + (X^T X)^{-1} X^T [A(X^T X)^{-1} A^T]^{-1} (c - A\hat{\beta})
\]

Denote \( \mu = X\hat{\beta} \) and \( \mu_0 = X\hat{\beta}_0 \), then
\[
(Y - \mu)^T (\mu_0 - \mu) = Y^T [I - X(X^T X)^{-1} X^T] (X^T X)^{-1} A^T [A(X^T X)^{-1} A^T]^{-1} (c - A\beta)
\]

\[
= Y^T (I - X(X^T X)^{-1} X^T) X^T [A(X^T X)^{-1} A^T]^{-1} (c - A\beta)
\]

\[
= Y^T [I - X(X^T X)^{-1} X^T] (X^T X)^{-1} \cdot A^T [A(X^T X)^{-1} A^T]^{-1} (c - A\beta)
\]

\[
= Y^T X (X^T X)^{-1} - (X^T X)^{-1} X^T (X^T X)^{-1} (X^T X)(X^T X)^{-1} \cdot A^T [A(X^T X)^{-1} A^T]^{-1} (c - A\beta)
\]

\[
= Y^T \cdot [A(X^T X)^{-1} A^T]^{-1} (c - A\beta) = 0
\]

**Conclusion** for \( Y = \hat{\mu}_0 = (Y - \hat{\mu}) + (\hat{\mu}_0 - \hat{\mu}) \):

\[
||Y - \hat{\mu}_0||^2 = ||Y - \hat{\mu}||^2 + ||\hat{\mu}_0 - \hat{\mu}||^2
\]
Variance estimator under reduced model (2)

Conclusion from previous slide

\[ \|Y - \hat{\mu}_0\|^2 = \|Y - \hat{\mu}\|^2 + \|\hat{\mu}_0 - \hat{\mu}\|^2 \]

Interpretation (Pythagoras) \(\hat{\mu}_0\) and \(\hat{\mu}\) are projections of \(Y\) onto \(X\), \(\hat{\mu}\) being the orthogonal projection. Hence, \(\Delta \hat{\mu}_0 \hat{\mu} Y\) is a rectangular triangle.

So

\[ \hat{\sigma}^2 = \frac{1}{n} \|Y - \hat{\mu}_0\|^2 \]
\[ = \frac{1}{n} \|Y - \hat{\mu}\|^2 + \frac{1}{n} \|\hat{\mu}_0 - \hat{\mu}\|^2 \]
\[ = \hat{\sigma}^2 + \frac{1}{n} \|X(X^T X)^{-1}A^T [A(X^T X)^{-1}A^T]^{-1} (c - A\hat{\beta})\|^2 \]
\[ = \hat{\sigma}^2 + \frac{1}{n} (c - A\hat{\beta})^T [A(X^T X)^{-1}A^T]^{-1} [A(X^T X)^{-1}A^T]^{-1} (c - A\hat{\beta}) \]

Conclusion

\[ \hat{\sigma}^2 - \hat{\sigma}^2 = \frac{1}{n} (c - A\hat{\beta})^T [A(X^T X)^{-1}A^T]^{-1} (c - A\hat{\beta}) \]

The proof for the \(F\)-test

We have

1. \(\hat{\sigma}^2 = \frac{n-p}{n} S^2 = \frac{(Y - X\hat{\beta})^T (Y - X\hat{\beta})}{n}\)
2. \(\hat{\sigma}^2 - \hat{\sigma}^2 = (A\hat{\beta} - c)^T [A(X^T X)^{-1}A^T]^{-1} (A\hat{\beta} - c) / n\)

We have to prove that

1. \(n \hat{\sigma}^2 / \sigma^2 = (n-p) S^2 / \sigma^2 \sim \chi^2_{n-p}\) (done before)
2. under \(H_0:\)

\[ n(\hat{\sigma}^2 - \hat{\sigma}^2) / \sigma^2 = (A\hat{\beta} - c)^T [A(X^T X)^{-1}A^T]^{-1} (A\hat{\beta} - c) / \sigma^2 \sim \chi^2_q \]

3. \(S^2\) and \(\hat{\beta}\) are independent

Therefore, all functions of \(S^2\) are also independent from all functions of \(\hat{\beta}\). More precisely \(\hat{\sigma}^2\) and \(\hat{\sigma}^2 - \hat{\sigma}^2\) are mutually independent.

An equivalent \(F\)-test

The distribution of \(-2 \log \Lambda\) is only asymptotically \(\chi^2\). The expression of \(-2 \log \Lambda\) leads to logarithms of sample variances. Intuitively, a ratio of sample variances suggests the use of an \(F\)-test, which is exact, and without the need of taking a \(\log\).

\sim ANOVA

Is the “extra sum of squares” (the part of the variance that is explained by giving up the restrictions imposed by the null hypothesis) significant?

Formally: \(F = \frac{(\hat{\sigma}^2 - \hat{\sigma}^2) / q}{\hat{\sigma}^2 / (n-p)}\)

We have: \(F = \frac{n-p}{q} \frac{\hat{\sigma}^2 - \hat{\sigma}^2}{\hat{\sigma}^2} = \frac{n-p}{q} \left( \frac{n-p}{\hat{\sigma}^2} - 1 \right) = \frac{n-p}{q} (\Lambda - n/2 - 1)\)

We now show that \(F \sim F_{q,n-p}\)

The proof for the \(F\)-test: distribution under \(H_0\)

As \(\hat{\beta} \sim N(\beta, (X^T X)^{-1}\sigma^2)\)

we have that \(A\hat{\beta} - c\) is a vector with length \(q\), which is normally distributed with expected value \(A\beta - c = 0\) (under \(H_0\))

and with covariance matrix \(A(X^T X)^{-1}A^T \sigma^2\).

Then (property covariance matrices)

\((A\hat{\beta} - c)^T [A(X^T X)^{-1}A^T]^{-1} (A\hat{\beta} - c) / \sigma^2 \sim \chi^2_q\)
The proof for the $F$-test: independence $S^2$ and $\hat{\beta}$

We show that $(Y - X\hat{\beta})$ is independent from $\hat{\beta}$. As both are normally distributed, it suffices to show that they are uncorrelated.

The covariance matrix. Mind the positions of the transposes (outer product):

$$\text{cov}(Y - X\hat{\beta}, \hat{\beta}) = E[(Y - X\hat{\beta})\hat{\beta}^T] - E(Y - X\hat{\beta}) \cdot E\hat{\beta}^T$$

$$= E[(Y - X(X^TX)^{-1}X^TY) \cdot ((X^TX)^{-1}X^TY)^T - 0 \cdot \beta^T$$

$$= (I - X(X^TX)^{-1}X^T) \cdot E(YY^T) \cdot (X^TX)^{-1}X^T$$

$$= (I - X(X^TX)^{-1}X^T) \cdot (\sigma^2 I + X\beta\beta^TX^T)X(X^TX)^{-1}$$

$$= (I - X(X^TX)^{-1}X^T) \cdot (X^TX)^{-1} \cdot (\sigma^2 + X\beta\beta^T) = 0$$

Simultaneous confidence

We know that $\hat{\beta}$ and $S^2$ are independent, so it follows immediately that

$$\frac{(\hat{\beta} - \beta)^T (X^TX)(\hat{\beta} - \beta)}{S^2} \sim F_{p,n-p}$$

(This is the $F$-statistic for the full model, so in the test $H_0$: all $\beta_i$ zero, i.e., $q = p$)

So

$$\left\{ \beta : \frac{(\hat{\beta} - \beta)^T (X^TX)(\hat{\beta} - \beta)}{S^2} \leq pF_{p,n-p,\alpha} \right\}$$

is an $1 - \alpha$ confidence region for $\beta$

Testing a single linear combination of $\beta$

$$a^T\hat{\beta} \sim N(a^T\beta, a^T(X^TX)^{-1}a\sigma^2)$$

So

$$\frac{a^T\hat{\beta} - a^T\beta}{S\sqrt{a^T(X^TX)^{-1}a}} \sim t_{n-p}$$

So

$$P \left( \frac{|a^T(\hat{\beta} - \beta)|}{S\sqrt{a^T(X^TX)^{-1}a}} \geq t_{n-p,\alpha/2} \right) = \alpha$$

Alternative argument (for $H_0: a^T\beta = 0$)

Take $A = a^T$ on slide 43, and $c = 0$, then $q = 1$, and $F$ defined on slide 49

$$a^T(\hat{\beta} - \beta)^2 = \frac{S^2a^T(X^TX)^{-1}a}{a^T(X^TX)^{-1}a} \sim F_{1,n-p}$$

(under $H_0$)

It holds that for $B \in \{-1, 1\}$ independent from $F$:

$$F \sim F_{1,n-p} \iff B\sqrt{F} \sim t_{n-p}$$

Testing multiple linear combinations of $\beta$

We verify now that the simulateneous confidence interval follows indeed from the $F$ test for the full model.

We search for $M_\alpha$, so that $P \left( \max_a \frac{|a^T(\hat{\beta} - \beta)|}{S\sqrt{a^T(X^TX)^{-1}a}} \geq M_\alpha \right) = \alpha$

We square the expression and get

$$\left[ a^T(\hat{\beta} - \beta) \right]^2 = \left[ a^T(\hat{\beta} - \beta) \right]^T \left[ a^T(\hat{\beta} - \beta) \right] = (\hat{\beta} - \beta)^T(aa^T)(\hat{\beta} - \beta)$$

The value of the expression in $a$ is the same as in $ra$ with $r \in \mathbb{R}_0$, so we can maximise under the normalisation that $a^T(X^TX)^{-1}a = 1$

$X^TX$ is a $p \times p$ symmetric, full rank matrix and so it can be decomposed as

$X^TX = RR^T$ with $R$ a $p \times p$ invertible matrix. So, we can write that

$a^T(X^TX)^{-1}a = a^T(R^{-T}R^{-1})a = u^TR^T\hat{\beta} - \beta)u$ with $u = R^{-1}a$, so $a = Ru$.

We maximise over $u$,

$$\max_u (\hat{\beta} - \beta)^T R uu^T R^T (\hat{\beta} - \beta)$$

The numerator contains a squared inner product of $u$ with $R^T(\hat{\beta} - \beta)$.
According to Cauchy-Schwarz’s inequality:

\[(\hat{\beta} - \beta)^T Ruu^T R^T (\hat{\beta} - \beta) \leq \|u\|^2 \cdot \|R^T (\hat{\beta} - \beta)\|^2,\]

So \[\max_u (\hat{\beta} - \beta)^T Ruu^T R^T (\hat{\beta} - \beta) = \|R^T (\hat{\beta} - \beta)\|^2 = (\hat{\beta} - \beta)^T RR^T (\hat{\beta} - \beta)\]

Since \[(\hat{\beta} - \beta)^T RR^T (\hat{\beta} - \beta) \sim \text{\textit{pF}}_{p,n-p}\]

We can take \[M_\alpha = \sqrt{\text{\textit{pF}}_{p,n-p,\alpha}}\]

**Residuals and outliers**

**Residuals** (See also slides 6 and 12)

\[e = Y - \hat{\mu} = (I - P)Y\]

\[\Sigma_e = \sigma^2 (I - P)\]

This allows to standardise/studentise the residuals and then test for normality.

**Cook-distance; influence measure; influence points**

measures the influence of observation \(i\) on the eventual estimated vector:

\[D_i = \frac{(\hat{\beta}_{(i)} - \beta)^T (X^T X)(\hat{\beta}_{(i)} - \beta)/p}{S^2}\]

**Transformations**

**Model**

\[Y^{(\lambda)} = X\beta + \varepsilon\]

with \[Y^{(\lambda)}_i = \frac{Y^{\lambda - 1}_i}{\lambda}\]

**General: transformation of a random variable**

If \(Y = g(X)\) then

\[f_Y(y) = f_X(x(y))|J(y)| = f_X(g^{-1}(y))|J(y)|\]

with \[J = \text{det} \begin{bmatrix} \cdots & \cdots & \cdots \\ \cdots & \frac{\partial x_i}{\partial y_j} & \cdots \\ \cdots & \cdots & \cdots \end{bmatrix} \] and \[\frac{\partial x_i}{\partial y_j} = \frac{\partial y_i^{-1}(y)}{\partial y_j}\]

In our case \(Y^{(\lambda)} \sim N(X\beta, \sigma^2 I)\) takes the role of \(X\), so \[\frac{\partial x_i}{\partial y_j} = 0\] unless \(j = i\), in that case \[\frac{\partial x_i}{\partial y_j} = y_i^{\lambda - 1}\]

So \[J = \prod_{i=1}^{n} g_i^{\lambda - 1}\]

The likelihood becomes

\[L(\beta, \sigma^2, \lambda; Y) = \frac{1}{(2\pi\hat{\sigma}^2)^{n/2}} e^{-\frac{1}{2\hat{\sigma}^2} (Y^{(\lambda)} - X\hat{\beta})^T (Y^{(\lambda)} - X\hat{\beta})} \prod_{i=1}^{n} Y_i^{\lambda - 1}\]

Differentation for \(\sigma^2\) yields

\[\hat{\sigma}^2 = \frac{1}{n} (Y^{(\lambda)} - X\hat{\beta})^T (Y^{(\lambda)} - X\hat{\beta})\]

Plugging in

\[L(\beta, \hat{\sigma}^2, \lambda; Y) = \frac{1}{(2\pi\hat{\sigma}^{n/2})^2} e^{-n/2} \prod_{i=1}^{n} Y_i^{\lambda - 1}\]

This function can be evaluated for different values of \(\lambda\) (note that \(\hat{\sigma}^2\) in its turn is also a function of \(\lambda\))
PART 2: VARIABLE AND MODEL SELECTION

2.1 Motivation and setup p.61
2.2 Selection criteria p.70
2.3 Minimisation of the Information criterion p.??

Motivation (1): full or large models are not the best

The analysis until now assumed that the full (global) model is correct. For instance, the expressions $E\hat{\beta} = \beta$ and $E\hat{\beta} = \sigma^2 (X^TX)^{-1}$ are true only if indeed $Y$ can be written as $X\beta$ and no explanatory variable is missing.

In order to check if a specific covariate $x_i$ should be considered as a candidate in the global model, one could “play safe” put it in the full model and let the hypothesis test $H_0 : \beta_i = 0$ decide whether or not it is significant.

But, if we have to play safe for a very wide range of possible explanatory variables, a lot of parameters need to be estimated. This has two effects:
1. The estimators have larger standard errors
2. Few degree of freedom left for estimation of variance $\sigma^2$

So: larger errors AND more difficult to assess the uncertainty: hypothesis tests will have weak power, and really significant parameters may be left undiscovered.

Occam’s razor: Keep the model as simple as possible, but no simpler.

Motivation (2): the true model is not always the best model

Previous slide: the full/largest model is not the best to work in

This slide: the true/physical model may not be the best statistical model

Example: assume that two quantities, $x$ and $y$, are related by the physical law $y = \beta_0 + \beta_1 x + \beta_2 x^2$

We observe $Y_i = y_i + \epsilon_i$ for $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2$

• Estimate $y_0 = \beta_0 + \beta_1 x_0 + \beta_2 x_0^2$ for a given $x_0$

If $\beta_2$ is (very) small, the model $y_i = \beta_0 + \beta_1 x_i$ is incorrect, introduces bias but an estimator $\hat{\beta}_2$ may add more variance than it reduces bias

• Estimate $\beta_2$ or $\beta_2/\beta_1$, then we need an estimator $\hat{\beta}_2$

Conclusions
1. “Best” statistical model may not be the true, physical model
2. “Best” statistical model depends on “what you want to estimate”. This is called the focus
Motivation (3): the larger the model, the more false positives

Large models require more multiple testing issues

Objective and situation of model selection

- A procedure to limit the size of the full model
- **before** estimation + testing
- Compare nested models but also totally different models
- ↔ testing: no null hypothesis at input, and no statistical evidence at output
- Testing: asymmetry in $H_0 \leftrightarrow H_1$; model selection is symmetric between the candidate models
- Compare model selection $\sim$ forensic profiling: profiling is not a proof of guilt.
  After profiling, suspects are innocent until proven guilty ($= H_0$, presumption of innocence $\rightarrow$ hypothesis testing)
- **Information criterion** expresses how well the data can be described in a given model

Model selection and variable selection

**Variable selection**

- Selects covariates
- Covariates require parameter $\beta_i$ to be estimated

**Model selection**

- selects statistical model, given the covariates
  - polynomial, exponential, periodic, …
  - How many terms (degrees, frequencies) (variables-within-model selection)
- Model for distribution of observations
  - Example: exponential distribution versus Weibull
  - Complicated models introduce parameters not linked to covariates
- Model for interactions between covariates

"All models are wrong, but some are useful"

George Box, Journal of the American Statistical Association, 1976:

*Since all models are wrong the scientist cannot obtain a "correct" one by excessive elaboration. On the contrary following William of Occam he should seek an economical description of natural phenomena. Just as the ability to devise simple but evocative models is the signature of the great scientist so overelaboration and overparameterization is often the mark of mediocrity.*

**Conclusion:** Full/large models are not the best

- Large standard errors in estimators
- Few degrees of freedom to estimate noise variance
- As a result of these two: hypothesis tests with low power
- More false positives
- Even if true model is large...
The use of likelihood in model selection

We know likelihood from statistical inference where it is used for

- parameter estimation within a given model by MLE, maximum likelihood
- Hypothesis testing, e.g.: $\chi^2$ test for goodness of fit
- Validation of the model with significant parameters after testing (e.g.: $R^2$ statistics)

We want to use likelihood in model selection, i.e., before statistical inference to compare different models

---

2.2 Model selection methods and criteria

- Adjusted $R^2$: slides 71 and further on
- Prediction error and Mallows's $C_p$: slides 75 and further on
- Kullback-Leibler distance and Akaike's Information Criterion: slides 89 and further on
- The Bayesian Information Criterion or Schwarz Criterion: slides 100 and further on
- Generalised Cross Validation: a criterion mainly used in sparse variable selection; see slide 122

---

The use of likelihood in model selection (see also p.90)

- **Working principle**: maximize expected likelihood within each model, then select model by maximisation
- Expectation is defined in terms of (unknown) data generating process (DGP) = true/physical model (see also 63)
- Likelihood measures closeness of fit
- Problem: large models lead to better fit, but more noisy parameter estimators (this is exactly one of the arguments for model selection)
- Solution: likelihood has to be adjusted or penalized: closeness-complexity trade-off (where likelihood = closeness)
- Closeness-complexity trade-off is equivalent to bias-variance trade-off: both trade-offs measure expected distance to true model

---

Example of a validation of a given model after inference: $R^2$

Least squares is an orthogonal projection: with $\hat{\mu} = X \hat{\beta}$ we have $(Y - \hat{\mu})^T Y = 0$ and also $(Y - \hat{\mu})^T 1 = 0$ since $1$ is the first column of $X$ and the residual is orthogonal to all columns of $X$ So $\sum_{i=1}^n (Y_i - \hat{\mu}_i)(\hat{\mu}_i - Y) = 0$

$$\sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n (Y_i - \hat{\mu}_i)^2 + \sum_{i=1}^n (\hat{\mu}_i - Y)^2$$

with

**Sum of squares Total** $SS_T = \sum_{i=1}^n (Y_i - \bar{Y})^2$

**Sum of squares Regression** $SS_R = \sum_{i=1}^n (\hat{\mu}_i - Y)^2$

**Sum of squares Errors of Prediction or Sum of squared residuals** $SS_E = \sum_{i=1}^n (Y_i - \hat{\mu}_i)^2 = \sum_{i=1}^n e_i^2$ where (see also p.6,12) $e_i = Y_i - \hat{\mu}_i$

We define **Coefficient of determination** $R^2 = \frac{SS_R}{SS_T}$ $R^2 = 1 - \frac{SS_E}{SS_T}$
Using $R^2$ before inference

$R^2$ is a measure of how well the model explains the variance of the response variables. However, a submodel of a full model has always a smaller $R^2$, even if the full model only adds unnecessary, insignificant parameters to the submodel.

Therefore $R^2$ cannot be used before inference: it would just promote large models.

**Note** The denominator in $R^2$ is the same for all models. The numerator is (up to a constant) the log-likelihood of a normal model with known variance. So $R^2$ measures the likelihood of a model. The more parameters the model has, the more likely the observations become, because there are more degrees of freedom to make the observations fit into the model.

**Adjusted coefficient of determination**

$$\overline{R}^2 = 1 - \frac{SS_E/(n-p)}{SS_T/(n-1)}$$

Or

$$1 - \overline{R}^2 = \frac{n-1}{n-p}(1 - R^2)$$

The criterion that decides which model has higher quality is thus equivalent to a “hypothesis test” with critical value $F = 1$, instead of $F_{q,n-p,q}$. This illustrates the symmetry of the model selection criterion, compared to the testing procedure, which uses null and alternative hypotheses, with clearly unequal roles.

Adjusted $R^2$: interpretation

**Adjusted $R^2$: interpretation**

We know:

$$S^2 = \frac{(Y-\hat{\mu})^T(Y-\hat{\mu})}{n-p} = \frac{1}{n-p} \sum_{i=1}^n (Y_i - \hat{\mu})^2 = \frac{1}{n-p} SS_E$$

and

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{\mu})^2 = \frac{n-p}{n} S^2 = \frac{1}{n-p} SS_E = \frac{1}{n}(SS_T - SS_R) = \frac{1}{n}SS_T(1-R^2)$$

Suppose we have two nested models, i.e., one model containing all covariates of the other (this is: a full model with $p$ covariates and a reduced model with $q$ restrictions, so with $p-q$ out of $p$ free covariates), then the $F$-value in testing the significance of the full model is then:

$$F = \frac{(\hat{\sigma}_e^2/\hat{\sigma}_0^2)/(n-p)}{SS_T/(n-1)}$$

Dividing numerator and denominator by $SS_T$ leads to:

$$F = \frac{(R_0^2 - R_q^2)/(n-p)}{(1-R_q^2)(1-R_q^2)/(n-1)}$$

We rewrite this in terms of “adjusted $R^2$-

$$F = \frac{n-p}{q} \left[ \frac{1-R_q^2}{1-R_p^2} - 1 \right] + \frac{1-R_q^2}{1-R_p^2}$$

from which it follows that $\overline{R}^2_{p-q} \leq \overline{R}^2_p \Leftrightarrow F \geq 1$

Mallows’s $C_p$ statistic: prediction error

**Criterion** Choose the model with smallest mean prediction error

The prediction for covariate values $x_p$: $\hat{\mu}_p = x_p^T \tilde{\beta}_p = x_p^T (X_p^T X_p)^{-1} X_p^T Y$

where $X_p$ is a model with $p$ covariates ($p$ columns, $n$ rows of observations)

$x_p^T$ contains the covariate values in one point of interest. $x_p^T$ may or may not coincide with one of the rows in $X_p$.

The prediction error is the expected difference $E(\hat{\mu}_p - \mu)^2$ between the prediction and the true, error-free value of the response for the given covariate values.

We want to assess the quality of the model $X_p$. The precise covariate values in which we measure the quality are free in theory. However, since we observe the model in a limited number of covariate values, we restrict to those values. Indeed, it is hard in practice to say something about the quality of a prediction of a response value, if we even don’t observe that response value. In the points of observations, we still don’t know the exact (noise-free) response.
The bias-variance trade-off

**Prediction error = variance + bias^2**

\[ E(\hat{\mu}_p - \mu)^2 = \text{var}(\hat{\mu}_p) + (E\hat{\mu}_p - \mu)^2 \]

with \( \mu = x^T \beta \) in the correct model

Intuitively: the smaller the model, the smaller the prediction variance, but the larger the bias.

Because of linearity of expectations

\[ E\hat{\mu}_p = E(x_p^T(X_p^T X_p)^{-1} X_p^T Y) = x_p^T(X_p^T X_p)^{-1} X_p^T \mu \]

And \( \text{var}(\hat{\mu}_p) = x_p^T \Sigma_\beta x_p = \sigma^2 x_p^T (X_p^T X_p)^{-1} x_p \)

We search for the minimum average prediction error, where the average is taken over all points of observation, so we want to minimize

\[ \text{PE}(\hat{\beta}_p) = \frac{1}{n} \sum_{i=1}^{n} E(\hat{\mu}_{pi} - \mu_i)^2 \]

The prediction variance of a model

The average variance can we written as

\[ \frac{1}{\sigma^2} \sum_{i=1}^{n} \text{var}(\hat{\mu}_{pi}) = \sum_{i=1}^{n} x_{pi}^T (X_p^T X_p)^{-1} x_{pi} \]

\[ = \sum_{i=1}^{n} \text{Tr}(x_{pi}^T (X_p^T X_p)^{-1} x_{pi}) = \sum_{i=1}^{n} \text{Tr}((X_p^T X_p)^{-1} (x_{pi} x_{pi}^T)) \]

\[ = \text{Tr}(I_p) = p \]

thereby using \( X_p = [x_{pi}^T] \) so \( X_p^T X_p = \sum_{i=1}^{n} (x_{pi} x_{pi}^T) \)

The average variance is thus only dependent on the size of the model. It does not depend on the exact locations of the points in which the prediction quality is evaluated.

**Remark** if we assume that the working model is correct (i.e., contains the true model), then result on slide 30 applies: bias is zero

The average mean or expected error (bias)

\[ \frac{1}{\sigma^2} \sum_{i=1}^{n} (E\hat{\mu}_{pi} - \mu_i)^2 = \frac{1}{\sigma^2} (E\hat{\mu}_p - \mu)^T (E\hat{\mu}_p - \mu) \]

Now \( E\hat{\mu}_p = \left[ \begin{array}{c} x_p^T (X_p^T X_p)^{-1} X_p^T \\ \vdots \end{array} \right] = X_p (X_p^T X_p)^{-1} X_p \mu = P_p \mu \)

so the average bias is \( \frac{1}{n} \mu^T (I - P)^T (I - P) \mu = \frac{1}{n} \mu^T \Sigma_\beta \mu \)

This contains the unknown, error-free response variable. We try to estimate the bias by plugging in the observed values instead. We find the expression \( Y^T (I - P) Y = \text{SSE}_p \).

The expression of Mallows’ Cp

The estimator of the bias is biased again, as follows from

\[ E(Y^T (I - P) Y) = E[(\mu + \varepsilon)^T (I - P)(\mu + \varepsilon)] \]

\[ = \mu^T (I - P) \mu + E[\varepsilon^T (I - P) \mu] + E[\mu^T (I - P) \varepsilon] + E[\varepsilon^T (I - P) \varepsilon] \]

\[ = \mu^T (I - P) \mu + 0 + 0 + \text{Tr}(I - P) \sigma^2 = \mu^T (I - P) \mu + (n - p) \sigma^2 \]

The bias of the estimator of the bias, \((n - p) \sigma^2\), however only depends on the model size, no longer on the unknown response.

So \( \text{PE}(\hat{\beta}_p) = \frac{1}{n} E[\text{SSE}_p + (2p - n) \sigma^2] \)

The random variable on the right hand side (between the expectation brackets) is observable and it is an unbiased estimator of the prediction quality measure in the left hand side of the expression.

In most practical situations \( \sigma^2 \) has to be estimated by \( S^2 \) and so the studentized value

\[ C_p = \frac{\text{SSE}_p}{S^2} + 2p - n \]

is a consistent estimator for \( \frac{n}{\sigma^2} \text{PE}(\hat{\beta}_p) \)
Remark on the estimation of $\sigma^2$

The variance of the errors did not appear in the definition of $PE(\hat{\beta}_p)$, it comes from the development of the expression.

We impose that the variance estimator must not depend on the selected model (unlike $\hat{\beta}_p$)

Therefore, $\sigma^2$ must be kept outside the optimisation of the likelihood within a given model, $S^2$ must not be model-dependent

---

The closeness-complexity trade-off

The $C_p$ criterion reformulates the bias-variance trade-off on slide 76 as a trade-off between closeness ($SS_Ep$) and complexity ($2p$).

From $E(\hat{\mu}_p - y)^2 = var(\hat{\mu}_p) + (E\hat{\mu}_p - y)^2$ to $PE(\hat{\beta}_p) = \frac{1}{n}E[SS_Ep + (2p - n)p\sigma^2]$.

The closeness-complexity trade-off can also be found directly from the expression of the prediction error, see next slide.

---

Degrees of freedom

Mallows’s $C_p$ without going by variance and bias

$PE(\hat{\beta}_p) = \frac{1}{n}E[\parallel \hat{\mu}_p - \mu \parallel^2 + \frac{1}{n}\parallel \hat{\mu}_p - Y + Y - \mu \parallel^2] = \frac{1}{n}E[\parallel e_p + \varepsilon \parallel^2]$ 

$= \frac{1}{n}E[\parallel e_p + \varepsilon \parallel^2 + 2E(-e_T^\varepsilon) + E\parallel \varepsilon \parallel^2]$ 

$= \frac{1}{n}E[SS_E(\hat{\beta}_p)] + \frac{2}{n}E[(e - e_p)^T\varepsilon] - \frac{1}{n}E\parallel \varepsilon \parallel^2$ 

$= \frac{1}{n}E[SS_E(\hat{\beta}_p)] + \frac{2\nu_p}{n}\sigma^2 - \sigma^2$

where $\nu_p$ are the degrees of freedom, defined as $\nu_p = \frac{1}{\sigma^2}E[\varepsilon^T(\varepsilon - e_p)]$.

And define a variant of Mallows’s $C_p$ based on degrees of freedom and without studentisation:

$\Delta(\hat{\beta}_p) = \frac{1}{n}SS_E(\hat{\beta}_p) + \frac{2\nu_p}{n}\sigma^2 - \sigma^2$ 

Then $E[\Delta(\hat{\beta}_p)] = PE(\hat{\beta}_p)$

---

Degrees of freedom is size of the model

As the residual is given by $e_p = (I - P_p)Y$ (see slide 12)

For a given/fixed model we find $\nu_p = \frac{1}{\sigma^2}E[\varepsilon^T(\varepsilon - (I - P_p)(\mu + \varepsilon))] = \frac{1}{\sigma^2}E[\varepsilon^TP_p(e - (I - P_p)(\mu))]$ 

$= \frac{1}{\sigma^2}E[\varepsilon^TP_pe] - \frac{1}{\sigma^2}(E\varepsilon^T)(I - P_p)\mu = \frac{1}{\sigma^2}E[Tr(P_pe\varepsilon^T)] - 0$ 

$= Tr(P_p\sigma^2I) = p$
Degrees of freedom after optimisation

- For a least squares in a fixed model, we find \( \nu_p = p \) (see p.83).
- If \( X_p \) and \( P_p \) result from optimisation of Mallows's \( C_p \), the selection depends on the errors.
- \( \Delta(\hat{\beta}_p) \) (see definition p.82) is still unbiased, but \( \nu_p \neq p \), because the (selected) projection \( P_p \) depends on the errors \( \epsilon \). Errors can cause false positives.
- False positives are more likely when many true parameters are zero. Therefore, the effect of the optimisation is important in sparse models. Otherwise, it can be ignored. See further slide 121.

The use of \( C_p \)

- In the notation \( C_p \), the \( p \) refers to the size of the model under consideration. It suggests that \( C_p \) is a function of \( p \) only. But \( C_p \) can be used to compare two non-overlapping models with the same size.
- \( C_p \) measures the quality of the model in terms of prediction error. It does not measure the correctness of the model. 
  prediction \( \leftrightarrow \) estimation
- Nested models, such as polynomial regression where we include all powers up to \( p \): a larger model thus always includes a smaller one.
  Then \( C_p \) can be seen as a function of \( p \) The curve of \( C_p \) has a typical form with a minimum that is the optimal balance between bias (descending as function of \( p \)) and variance (proportional to \( p \)). If true model \( \in \{ \text{models} \} \), then bias \( \to 0 \).

Illustration of a nested model

Polynomial regression model \( Y_i = \sum_{k=0}^{p-1} \beta_k x_i^k + \epsilon_i \) with \( n = 30 \) observations in \([0, 1] \), and \( p = 8 \).
(Degree 7, specified in the simulation by 7 zeros \( \Rightarrow 6 \) extremes; two zeros are outsize \([0, 1] \))
Conclusions about the example

The true model is a polynomial of degree 7. It has two zeros outside the interval of observations. The presence of these zeros has little impact on the observations, it can hardly be detected. As a consequence, the minimum prediction error is reached for a simpler model: a polynomial of degree 5 ($p = 6$).

Question

1. Why is $PE(\tilde{\beta}_p)$ almost a linear function for $p$ large?

Expected log-likelihood in model selection

- See also slide 69

\[ \ell(\hat{\beta}_p, \hat{\sigma}_p^2) = ELL(\hat{\beta}_p, \hat{\sigma}_p^2; Y) = \frac{1}{n} \sum_{i=1}^{n} E \log f_{Y_i}(Y_i; \hat{\beta}_p, \hat{\sigma}_p^2) \]

**Definition**

\[ \ell(\hat{\beta}_p, \hat{\sigma}_p^2) = \frac{1}{n} \sum_{i=1}^{n} \int_{-\infty}^{\infty} \log f_{Y_i}(u; \hat{\beta}_p, \hat{\sigma}_p^2) g_{Y_i}(u; \sigma^2) du \]

**Estimation**

- Substitute $\hat{\beta}_p$ and $\hat{\sigma}_p^2$ by sample based estimators:
  \[ \ell(\hat{\beta}_p, \hat{\sigma}_p^2) \]
  is random

- Estimate expected value (integral) from sample $Y$:
  \[ \hat{\ell}(\hat{\beta}_p, \hat{\sigma}_p^2) = \frac{1}{n} \sum_{i=1}^{n} \log f_{Y_i}(Y_i; \hat{\beta}_p, \hat{\sigma}_p^2) \]

\[ E \left[ \hat{\ell}(\hat{\beta}_p, \hat{\sigma}_p^2) \right] = \ell(\hat{\beta}_p, \hat{\sigma}_p^2) \quad \text{but} \quad E \left[ \ell(\hat{\beta}_p, \hat{\sigma}_p^2) \right] \neq E \left[ \hat{\ell}(\hat{\beta}_p, \hat{\sigma}_p^2) \right] \]

Kullback-Leibler distance and expected log-likelihood

- Data generating density – data generating process – DGP $g_Y(y; \sigma)$: (joint density of observations $Y$)

- Models for selection $f_Y(y; \hat{\beta}_p, \hat{\sigma}_p^2)$: where $\hat{\beta}_p$ has $p$ parameters

- then Kullback–Leibler (KL) divergence

\[ KL \left( g, f(\cdot; \hat{\beta}_p, \hat{\sigma}_p^2) \right) = \frac{1}{n} \sum_{i=1}^{n} \left[ E \log g_{Y_i}(Y_i; \sigma^2) - E \log f_{Y_i}(Y_i; \hat{\beta}_p, \hat{\sigma}_p^2) \right] \]

measures distance true model – statistical model

- Again: bias–variance balance

\[ E \log g_{Y_i}(Y_i; \sigma^2) \]
does not depend on chosen model, so

\[ KL \left( g, f(\cdot; \hat{\beta}_p, \hat{\sigma}_p^2) \right) = \text{Constant} - ELL(\hat{\beta}_p, \hat{\sigma}_p^2; Y) \]

with $LL(\hat{\beta}_p, \hat{\sigma}_p^2; Y) = \frac{1}{n} \sum_{i=1}^{n} \log f_{Y_i}(Y_i; \hat{\beta}_p, \hat{\sigma}_p^2)$

\[ KL \text{-distance} = \text{constant} - \text{expected log-likelihood} \]

Estimating KL from a sample

**Problem**: using the same sample for

- parameter estimation within a model and
- distance of model to data generating density
Suppose linear model – normal errors

Let $\mu = E(Y)$ and $\mu_p = Xp\hat{\beta}_p$

$$\ell(\hat{\beta}_p, \hat{\sigma}^2_p) = -\frac{1}{n} \sum_{i=1}^{n} E \left[ \frac{(Y_i - \hat{\mu}_p)^2}{2\hat{\sigma}^2_p} + \frac{1}{2} \log(2\pi\hat{\sigma}^2_p) \right]$$

$$= -\frac{1}{n} \sum_{i=1}^{n} \left[ (\mu_i - \hat{\mu}_p)^2 + E(Y_i - \mu_i)^2 \frac{2\hat{\sigma}^2_p}{2\hat{\sigma}^2_p} + \frac{1}{2} \log(2\pi\hat{\sigma}^2_p) \right]$$

$$= -\frac{1}{n} \sum_{i=1}^{n} \left[ (\mu_i - \hat{\mu}_p)^2 + \frac{\sigma^2}{\hat{\sigma}^2_p} + \frac{1}{2} \log(2\pi\hat{\sigma}^2_p) \right]$$

Note $\hat{\sigma}^2_p \neq \sigma^2$: (model $\leftrightarrow$ data generating process)

$$\ell(\hat{\beta}_p, \hat{\sigma}^2_p) = -\frac{1}{n} \sum_{i=1}^{n} \left[ (\mu_i - \hat{\mu}_p)^2 + \frac{\sigma^2}{\hat{\sigma}^2_p} + \frac{1}{2} \log(2\pi\hat{\sigma}^2_p) \right]$$

Taking $\hat{\sigma}^2_p = \frac{1}{\nu_p} \sum_{i=1}^{n} (Y_i - \hat{\mu}_p)^2$ (where $\nu_p$ to be chosen) we have

$$\ell(\hat{\beta}_p, \hat{\sigma}^2_p) = -\frac{1}{\nu_p} \sum_{i=1}^{n} \left[ \frac{(Y_i - \hat{\mu}_p)^2}{2\hat{\sigma}^2_p} + \frac{1}{2} \log(2\pi\hat{\sigma}^2_p) \right] = -\frac{1}{\nu_p} \frac{1}{2} \frac{n}{n} - \frac{1}{2} \log(2\pi\hat{\sigma}^2_p)$$

Development of AIC for linear model (3)

$$E \left[ \ell(\hat{\beta}_p, \hat{\sigma}^2_p) \right] = -\frac{1}{n} \sum_{i=1}^{n} E \left[ \frac{(\mu_i - \mu_{pi})^2}{2\hat{\sigma}^2_p} + \frac{\sigma^2}{2\hat{\sigma}^2_p} + \frac{1}{2} \log(2\pi\hat{\sigma}^2_p) \right]$$

$$= -\frac{1}{2} E \left[ \frac{\sigma^2}{\hat{\sigma}^2_p} \cdot \left( \frac{1}{\nu_p} \hat{\sigma}_p^2 \frac{||\mu - \hat{\mu}_p||^2}{\hat{\sigma}_p^2 + 1} \right) + \log(2\pi\hat{\sigma}^2_p) \right]$$

$$= -\frac{1}{2} E \left( \frac{\sigma^2}{\hat{\sigma}^2_p} \right) E \left( \frac{1}{\nu_p} \hat{\sigma}_p^2 \cdot ||\mu - \hat{\mu}_p||^2 \right) + E \left( \log(2\pi\hat{\sigma}^2_p) \right)$$

We now use

1. $Y \sim \chi^2(\nu) \rightarrow E(1/Y) = 1/(\nu - 2)$
2. (cfr.slides 77 and 78) $E(\hat{\sigma}_p^2) = p + \mu^T(I - P)\mu$

Hence, if $\beta_p$ contains the true model, then

$$E \left[ \ell(\hat{\beta}_p, \hat{\sigma}^2_p) \right] = -\frac{1}{2} \left( \frac{\nu_p}{n - p - 2} \right) \left( \frac{p}{n} + 1 \right) - \frac{1}{2} E \left( \log(2\pi\hat{\sigma}^2_p) \right)$$

while (p.92) $E \left[ \ell(\hat{\beta}_p, \hat{\sigma}^2_p) \right] = -\frac{1}{2} \left( \frac{\nu_p}{n} \right) - \frac{1}{2} E \left( \log(2\pi\hat{\sigma}^2_p) \right)$

Akaike’s Information Criterion — AIC – definition

$$E \left[ \ell(\hat{\beta}_p, \hat{\sigma}^2_p) \right] = E \left[ \ell(\hat{\beta}_p, \hat{\sigma}^2_p) \right] + \frac{1}{\nu_p} \left( \frac{\nu_p}{n - p - 2} \right) \left( \frac{p}{n} + 1 \right)$$

$$= E \left[ \ell(\hat{\beta}_p, \hat{\sigma}^2_p) \right] - \frac{p + 1}{n} \left( \frac{\nu_p}{n - p - 2} \right)$$

As $\nu_p/(n - p - 2) = O(1)$, we can define AIC as

$$AIC(\hat{\beta}_p) = \frac{1}{n} \sum_{i=1}^{n} \log f_Y(Y_i; \hat{\beta}_p, \hat{\sigma}^2_p) - 2 \left( \frac{p + 1}{n} \right)$$

Intuition the penalty comes from the fact that max.likelihood estimator substituted into sample likelihood expression fails to detect the noise variance (second term on second line on p.92 disappears)
AIC – sketch of general proof

- Let $\hat{\theta}_p$ minimize $\ell(\theta_p) = \frac{1}{n} \log f_Y(y; \theta_p)$ (expected log-likelihood)
- Let $\theta_p^*$ minimize $\ell(\theta_p) = \frac{1}{E} \log f_Y(y; \theta_p)$ (sample log-likelihood)
- $\theta_p^*$ is the least false parameter value

If the DGP belongs to the model family, then $\theta_p^*$ is the true parameter value

Taylor series for $\nabla \ell(\theta)$ around $\theta_p^*$: $\nabla \ell(\hat{\theta}_p) \approx \nabla \ell(\theta_p^*) + H(\theta_p^*)(\hat{\theta}_p - \theta_p^*)$

By definition, $\nabla \ell(\hat{\theta}_p) = 0$, so the score $\nabla \ell(\theta_p^*)$ satisfies

$\nabla \ell(\theta_p^*) \approx -H(\theta_p^*)(\hat{\theta}_p - \theta_p^*)$ (similar for $H(\theta)$)

Argument for the latter approximation:

- The score can be written as $\nabla \ell(\theta_p^*) = \frac{1}{n} \sum_{i=1}^{n} \nabla \theta Y_i; \theta_p^*$
  This has the form of a sample mean, fluctuating around $E(\nabla \ell(\theta_p^*)) = \nabla \ell(\theta_p^*) = 0$
  hence, it can be expected that a standardised value converges in distribution to a
  multivariate standard normal. (CLT)
- The Hessian: $H(\theta_p^*) = \frac{1}{n} \sum_{i=1}^{n} H \log f_Y(Y_i; \theta_p^*)$ does not fluctuate around
  zero, but converges in probability to a constant
- Convergence rates: score $O_p(1/\sqrt{n})$; Hessian $O_p(1/n)$
- Conclusion: approximate Hessian by $H(\theta_p^*)$, keep score as a random vector

For the proof of AIC, we need some results under model misspecification

Taylor series for use in the proof of AIC

- $\nabla \ell(\hat{\theta}_p) = \ell(\theta_p^*) + \frac{1}{2} \theta_p^* \theta_p^* H(\theta_p^*)(\hat{\theta}_p - \theta_p^*)$
  (expected log-likelihood)

AIC - general proof (ctd.)

From slide 98, we have $\hat{\theta}_p - \theta_p^* \approx -H(\theta_p^*)^{-1} \nabla \ell(\theta_p^*)$

Substitute into the expression for $\ell(\hat{\theta}_p) - \ell(\theta_p^*)$ on slide 97:

$\ell(\hat{\theta}_p) - \ell(\theta_p^*) \approx \ell(\theta_p^*) - \ell(\theta_p^*) + \frac{1}{2} \theta_p^* \theta_p^* H(\theta_p^*)(\hat{\theta}_p - \theta_p^*)$

Expected values

Use: scalar = Tr(scalar) — Tr(AB) = Tr(BA) — $\nabla \ell(\theta_p^*)$ has zero mean

$E(\ell(\hat{\theta}_p)) - E(\ell(\theta_p^*)) \approx 0 - Tr[H(\theta_p^*)^{-1} E(\nabla \ell(\theta_p^*) \nabla \ell(\theta_p^*)^T)]$

$\approx -Tr[H(\theta_p^*)^{-1} \Sigma \nabla \ell(\theta_p^*)]$ If $\theta_p^*$ is true model (DGP), then $-H(\theta_p^*)^{-1} \Sigma \nabla \ell(\theta_p^*)$ (Fisher information matrix), so $E(\ell(\hat{\theta}_p)) - E(\ell(\theta_p^*)) \approx \text{Tr}(I_p/n) = p/n$

So, take $\text{AIC}(\hat{\theta}_p) = 2(\hat{\ell}(\theta_p)) - \frac{2p}{n}$
Bayesian Information Criterion (BIC) or Schwarz Criterion

The Bayesian framework

- No true model, all models have prior probability of being the DGP
- Let $S_m$ be the event that the data were generated in model $m$
- Let $\Theta_m$ be the set of possible values of $\theta$ within the $m$th model
- Let $\varphi(\theta|S_m)$ a prior density for the parameters in model $m$
- Then (Bayes) $P(S_m|y) = \frac{P(S_m)f_Y(y|S_m)}{f_Y(y)} = \frac{P(S_m)f_Y(y|S_m)}{\sum_{k=1}^{K}P(S_k)f_Y(y|S_k)}$
  with $f_Y(y|S_m) = \int_{\Theta_m} f_Y(y; \theta_m|S_m)\varphi(\theta|S_m)d\theta_m = \int_{\Theta_m} f_Y(y; \theta_m|S_m)\varphi(\theta|S_m)d\theta_m$
  - If $P(S_m)$ is constant, then maximisation of $P(S_m|y)$ amounts to maximisation of $f_Y|S_m(y)$

BIC - formula

Development of expression p.100, with $\hat{\theta}_m$, the max. likelihood estimator of $\theta_m$:

$f_Y|\theta_m(y; \theta_m) = \exp \left[ \sum_{i=1}^{n}\log f_Y(Y_i; \theta_m) \right] \approx \exp \left[ n\hat{\ell}(\hat{\theta}_m) - \frac{1}{2}(\theta_m - \hat{\theta}_m)^T nH\hat{\ell}(\hat{\theta}_m)(\theta_m - \hat{\theta}_m) \right]$

Then we can apply Laplace's approximation for an integral (see slide 105)

$f_Y(y|S_m) = \int_{\Theta_m} f_Y(y; \theta_m|S_m)\varphi(\theta|S_m)d\theta_m$

$\approx \exp \left[ \sum_{i=1}^{n}\log f_Y(Y_i; \theta_m) \right] \exp \left[ \frac{n}{2} \left( \hat{\ell}(\hat{\theta}_m) - \frac{1}{n}\log(\det(H\hat{\ell}(\hat{\theta}_m))) \right) \right]$

Then we have

$P(S_m|y) \approx \frac{P(S_m)exp[BIC(\hat{\theta}_m)/2]}{\sum_{k=1}^{K}P(S_k)exp[BIC(\hat{\theta}_k)/2]}$

Differences in setup AIC — BIC

AIC has the notion of a data generating process/true model

BIC has starts from the probability that the data were generated by model $m$, for all models under consideration

AIC: Every model $m$ has a true or least false parameter $\theta_m^*$

BIC: Within every model, the parameter values have random/prior distribution

AIC tries to find the model with best expected (log-)likelihood.

BIC tries to maximise the posterior probability, given the observations, in case of non-informative prior, this is maximum sample likelihood (see slide 102

BIC - formula simplified

All terms on slide 102 have the factor $\exp(n/2)$ in common. Between brackets, we see, for $n \to \infty$,

- in the first term $\hat{\ell}(\hat{\theta}_m) = \frac{1}{n}\sum_{i=1}^{n}\log f_Y(Y_i; \hat{\theta}_m) = O_p(1)$ (see p.90)
- in the second term $O\log(n)/n$
- in all other terms $O(1/n)$ or $O_p(1/n)$
  (In particular, the Hessian converges to constant (independent from $n$) in probability, see p.98)

Defining

$BIC(\hat{\theta}_m) = 2\hat{\ell}(\hat{\theta}_m) - \log(n) \cdot \frac{p}{n} = \frac{2}{n}\sum_{i=1}^{n}\log f_Y(Y_i; \hat{\theta}_m) - \log(n) \cdot \frac{p}{n}$

we have

$P(S_m|y) \approx \frac{P(S_m)exp[BIC(\hat{\theta}_m)/2]}{\sum_{k=1}^{K}P(S_k)exp[BIC(\hat{\theta}_k)/2]}$
A classical application of Laplace’s approximation: Stirling’s formula

\[ n! \approx \sqrt{2\pi n}(n/e)^n \]

\[ \Gamma(z+1) \approx \sqrt{2\pi} e^{-z^2/2} \]

\[ \Gamma(z+1) = \int_0^\infty e^{-z^2} dx = \int_0^\infty e^{-zu} u^z du = z^{z+1} \int_0^\infty e^{-2zu} du = z^{z+1} \int_0^\infty e^{z[\log(u) - u]} du \]

The function \( g_z(u) = z[\log(u) - u] \) has a global minimum in \( u = 1 \), with \( g_z'(1) = 0 \) and \( g_z''(u) = -z/u^2 \), so

\[ \Gamma(z+1) \approx z^{z+1} \sqrt{2\pi} e^{z\log(z/e)} \frac{1}{\sqrt{|g_z''(1)|}} = \sqrt{2\pi} \frac{1}{\sqrt{z/e}} z^z \]

**BIC as a Bayes factor**

When all prior probabilities are equal (or unknown, and thus supposed to be equal, for convenience), the ratio of conditional distributions controls the selection.

This ratio is known as the **Bayes factor**

\[ \text{Bayes factor} = \frac{f_Y(y|S_m)}{f_Y(y|S_k)} = \exp \left[ \frac{\text{BIC}({\hat{\theta}_m})}{2} - \frac{\text{BIC}({\hat{\theta}_k})}{2} \right] \]

Bayes factors are the Bayesian analogues for likelihood ratios

**Laplace’s approximation for integrals**

Let \( I = \int_{\mathbb{R}^n} f(x)e^{g(x)} dx \), and let \( x_0 \) be the global maximum of \( g(x) \), then

\[ g(x) \approx g(x_0) + \frac{1}{2}(x-x_0)^T Hg(x_0)(x-x_0) = g(x_0) - \frac{1}{2}(x-x_0)^T [Hg(x_0)](x-x_0) \]

Moreover \( f(x) \approx f(x_0) + \nabla^T f(x_0)(x-x_0) + \frac{1}{2}(x-x_0)^T Hf(x_0)(x-x_0) \)

So

\[ I \approx e^{g(x_0)} \int_{\mathbb{R}^n} \exp \left[ -\frac{1}{2}(x-x_0)^T [Hg(x_0)](x-x_0) \right] (f(x_0) + \nabla^T f(x_0)(x-x_0) + \ldots) dx \]

The exponential factor can be read as the density of a multivariate normal around \( x_0 \) times the missing constant \( (2\pi)^{n/2} |\text{det} (Hg(x_0^{-1})^{1/2}) \), hence

\[ \int_{\mathbb{R}^n} \exp \left[ -\frac{1}{2}(x-x_0)^T [Hg(x_0)](x-x_0) \right] \nabla^T f(x_0)(x-x_0)dx = 0 \]

\[ I \approx e^{g(x_0)} f(x_0)(2\pi)^{n/2} |\text{det} (Hg(x_0))|^{-1/2} \]

**Algorithms for searching the best model**

- Suppose the full model has \( K \) possible variables, then the number of possible models in \( 2^K \)
- Computing all adjusted \( R^2 \)'s, \( C_p \) scores or other criterion is impossible (exponential computational complexity)
- Possible strategies:
  1. **Forward selection**: Start with a zero model, then gradually add the most significant parameter in a test with \( H_0 \) the current model.
  2. **Backward elimination**
  3. **Forward-backward alternation**: after a variable has been added, it is tested if any variable in the model can now be eliminated without significant increase in RSS.
  4. **Grouped/structured selection**: the routine may impose that some parameters must be included before others. For instance, no interaction terms without main effects. Imposing nested models is an extreme example.
ART 3: HIGH DIMENSIONAL MODEL SELECTION

Sparsity

- **Parsimonious** model selection: we search for models where “some” covariates have exactly zero contribution.
- **Sparse** model selection: the number of zeros dominates (by far) the number of nonzeros

Sparsity will be used as an assumption in high dimensional model selection, thereby defining a solution. On the other hand, sparsity will cause problems w.r.t. false positives.

Regularisation: from combinatorial to convex constraints

- In the constraint optimization problem \( \text{RegLS}_\ell(\hat{\beta}) = \|Y - X \cdot \hat{\beta}\|^2 + \lambda_\ell \), the constraint or penalty equals \( p = \sum_i \beta_i^0 = \|\beta\|_0 \), where we take \( \ell_0 \) regularization = combinatorial integer programming problem.
- Replace \( \ell_0 \) by \( \ell_1 \): \( \rightarrow \text{convex} \) quadratic programming problem

<table>
<thead>
<tr>
<th>Combinatorial constraint</th>
<th>Algebraic constraint</th>
<th>Convex constraint</th>
<th>Linear constraint</th>
</tr>
</thead>
<tbody>
<tr>
<td>( X \beta = Y )</td>
<td>( X \hat{\beta} = Y )</td>
<td>( X \hat{\beta} = Y )</td>
<td>( X \hat{\beta} = Y )</td>
</tr>
<tr>
<td>( \sum_i</td>
<td>\beta_i</td>
<td>&gt; 0 )</td>
<td>( \sum_i \beta_i^0 &gt; 0 )</td>
</tr>
<tr>
<td>( \sum_i \beta_i^0 = 1 )</td>
<td>( \sum_i</td>
<td>\beta_i</td>
<td>= 1 )</td>
</tr>
<tr>
<td>( \sum_i</td>
<td>\beta_i</td>
<td>^2 = 1 )</td>
<td>( \sum_i \beta_i^0 = 1 )</td>
</tr>
</tbody>
</table>

Equivalence \( \lambda \leftrightarrow p \)

- For \( q < 2 \), the outcome of the regularized minimum sum of squared residuals is a sparse model, with, say, \( p \) selected covariates: with each \( \lambda \) corresponds one \( p \).
- So, we can also regularize with \( p \)
- The outcomes are models that are not (necessarily) nested, but they can be ordered (put on a line) from simple to complex.
- We can thus, for instance, plot \( \text{PE}(\hat{\beta}_p) \) as a function of \( p \), just as in the case with nested models.

Model selection as a constraint optimization problem

Constraint optimization problem = regularized least squares

\[ \text{RegLS}(\hat{\beta}) = \|Y - X \cdot \hat{\beta}\|^2 + \lambda_\ell \]

- Minimum over \( \hat{\beta} \)
-Defines the estimator within a model
-Operates within full model, but estimator is parsimonious (if \( q < 2 \))

\( \text{RegLS}(\hat{\beta}) \) depends on smoothing parameter \( \lambda \). Parameter can be estimated using penalized likelihood, for instance Mallows’ \( C_p \):

\[ C_p = \frac{\text{SSR}_p}{n} + 2p - n \]

- Information criterion
- Minimize over \( \lambda \) (or equivalently: over model size \( p \))
- Smoothing parameter \( \lambda \) determines model size \( p \) (and vice versa)
Solving the $\ell_1$ constraint problem

Steepest descend for quadratic forms

Suppose $F(\beta) = \|Y - X\beta\|^2_2 = \sum_{i=1}^n \left(Y_i - \sum_j X_{ij}\beta_j\right)^2$

then $\frac{\partial F}{\partial \beta_k} = \sum_{i=1}^n 2 \left(Y_i - \sum_j X_{ij}\beta_j\right) X_{ik}$

Sum over $i$ is an inner product of the vectors $Y - X\beta$ and the $k$th column of $X$, so all $k$ together, the gradient vector becomes $X^T(Y - X\beta)$

Karush-Kuhn-Tucker conditions

If $\hat{\beta}$ solves the $\ell_1$ constraint optimization problem

$\text{RegLS}_{\ell_1}(\beta) = \|Y - X\beta\|^2_2 + 2\lambda \|\beta\|_1$

then

$X_j^T(Y - X\hat{\beta}) = \lambda \text{sign}(\hat{\beta}_j)$ if $\hat{\beta}_j \neq 0$

$|X_j^T(Y - X\hat{\beta})| < \lambda$ if $\hat{\beta}_j = 0$

Intuition

• $\beta_j$ in the model
  If $\hat{\beta}_j \neq 0$, then $\frac{\partial \text{RegLS}_{\ell_1}(\beta)}{\partial \beta_j}(\hat{\beta})$ is continuous, and it must be zero. That is expressed by the first line in the Karush-Kuhn-Tucker conditions

• $\beta_j$ not in the model
  If $\hat{\beta}_j = 0$, then the partial derivative is discontinuous. The steepest decrease of $\|Y - X\beta\|^2_2$ by letting $\hat{\beta}_j$ deviate from zero, should then be smaller than the increase in the penalty, i.e., the cost (price to pay) for introducing $\beta_j$ into the model is higher than the gain:

$\text{gain} < \text{cost} \Leftrightarrow \frac{\partial}{\partial \hat{\beta}_j} \|Y - X\beta\|^2_2 < 2\lambda \frac{\partial}{\partial \beta_j} \|\beta\|_1 \Leftrightarrow |X_j^T(Y - X\hat{\beta})| < \lambda$

Interpretation of the Karush-Kuhn-Tucker conditions

The solution of the $\ell_1$ constraint problem will be sparse, but it does not coincide with the $\ell_0$ constraint solution. Indeed, the solution that satisfies the KKT conditions cannot be an orthogonal projection (a least squares solution).

On the next slides is a simple, yet important, special case: soft- versus hard-thresholding. The latter is a $\ell_0$, and thus a projection, the former is projection plus shrinkage

Then, on slide 117, we extend the interpretation to the general case.

Soft- and Hard-Thresholding

• Suppose $X = I$, so we have the observational model $Y = \beta + \varepsilon$
  In this model, both $\ell_0$ and $\ell_1$-penalized closeness-of-fit can be optimized componentwise

• $\ell_0$-penalized closeness-of-fit:

$\text{RegLS}_{\ell_0}(\beta) = \sum_{i=1}^n (Y_i - \beta_i)^2 + \lambda^2 I(|\beta_i| > 0)$

Solution is **Hard-thresholding**: $\hat{\beta}_i = \text{HT}_\lambda(Y_i)$, where $\text{HT}_\lambda(x)$ is the following function

• $\ell_1$-penalized closeness-of-fit:
RegLS_{ℓ_1}(β) = \sum_{i=1}^{n}(Y_i - β_i)^2 + 2\lambda|β_i|

Solution is **Soft-thresholding**: \( \hat{β}_i = \text{ST}_λ(Y_i) \), where \( \text{ST}_λ(x) \) is the following function

Shrinkage in \( ℓ_1 \) constraint least squares for \( Y = Xβ + ε \)

Let \( I \) be the set of selected components, then the Karush-Kuhn-Tucker conditions are equivalent to \( X_I^TX_I\hat{β}_I = \text{ST}_λ(X_I^TY) \)

From which we conclude that within the optimal selection, the estimation is not least squares, but it contains a shrinkage

Note that \( X^TX\hat{β} \neq \text{ST}_λ(X^TY) \) (i.e., without selection \( I \))

**LASSO/Basis pursuit**

- \( ℓ_1 \) penalization/regularization/constraint is called least absolute shrinkage and selection operator (LASSO) or Basis Pursuit
- For given \( λ \), \( ℓ_1 \) penalization leads to the same degree of sparsity as \( ℓ_0 \) (see Figure on slide 103)
- For fixed \( λ \), and if all nonzero \( β \) are large enough, \( ℓ_1 \) penalization is variable selection consistent: for \( n \to \infty \), the set of nonzero variables in the selection equals the true set with probability tending to one
- The convex optimization problem can be solved by quadratic/dynamic programming or by specific methods, such as
  - Least Angle Regression (LARS), a direct solver
  - Iterative Soft Thresholding (It. ST), an iterative solver
- We discuss both on the subsequent slides

**Least Angle Regression**

- Brings in a new variable that with maximum projection of current residual, i.e., component with highest magnitude in the vector \( X^T(Y - X\hat{β}_0) \)
- Moves along equiangular vector \( u_I : \hat{β}_I = \hat{β}_0 + αu_I \) with
  \[
  u_I = X_I(X_I^TX_I)^{-1}1_I\sqrt{1_I^TX_I(X_I^TX_I)^{-1}1_I}
  \]
  where \( I \) is set of active variables (variables currently in the model), until one variable not in \( I \) has the same projection of the new residual \( X^T(Y - X\hat{β}_I) \)
- Stopping criterion: Mallows’s \( C_p \): we stop if (we think that) \( C_p(p) \) has reached a minimum
Iterative soft-thresholding

- General $X$, observational model $Y = X\beta + \varepsilon$
- We want $\|Y - X\hat{\beta}\|_2^2$ as small as possible, under the constraint that $\|\hat{\beta}\|_1 = \sum_i |\beta_i|$ is restricted
- Suppose we have an estimator $\hat{\beta}(r)$, then we improve this estimator in two steps
  - We proceed in the direction of the steepest descend in $\|Y - X\hat{\beta}\|_2^2$.
    The steepest descend is $X^T (Y - X\hat{\beta}(r))$
    We define $\hat{\beta}(r+1/2) = \hat{\beta}(r) + X^T (Y - X\hat{\beta}(r))$
  - We search for $\hat{\beta}(r+1)$ which is as close as possible to $\hat{\beta}(r+1/2)$, but which has restricted $\ell_1$-norm, so, we take $\hat{\beta}(r+1) = ST_\lambda(\hat{\beta}(r+1/2))$

Degrees of freedom in sparse variable selection

Degrees of freedom (see definition slide 82) $\nu_p = \frac{1}{\sigma^2} \mathbb{E}[\varepsilon^T (\varepsilon - e_p)]$

Generalized Cross Validation (GCV): compromise between CV and Cp

Definition $GCV(\hat{\beta}_p) = \frac{\frac{1}{n} \mathbb{E}[\varepsilon^T (\hat{\beta}_p) - e_p]}{(1 - \frac{\nu_p}{n})^2}$

This comes from a formalization + simplification of the routine of cross validation

GCV is an estimator of the prediction error (like Cp)

The use of GCV assumes sparse models

GCV combines benefits CV and Cp
- Like Cp, GCV is much faster than CV
- Like CV, GCV does not need a variance estimator

Implicit variance estimator is quite robust

Generalised Cross Validation

GCV also estimates the prediction error, however without having to estimate the variance of the errors.

$GCV(\hat{\beta}_p) = \frac{\frac{1}{n} \mathbb{E}[\varepsilon^T (\hat{\beta}_p)]}{(1 - \frac{\nu_p}{n})^2}$ while $\Delta(\hat{\beta}_p) = \frac{1}{n} \mathbb{E}[\varepsilon^T (\hat{\beta}_p) + \frac{2\nu_p}{n} \sigma^2 - \sigma^2]$ (see slide 82 for definition of $\Delta(\hat{\beta}_p)$)

Working GCV is based on

$GCV(\hat{\beta}_p) - \sigma^2 = \frac{\Delta(\hat{\beta}_p) - (\frac{\nu_p}{n})^2 \sigma^2}{(1 - \frac{\nu_p}{n})^2}$

$\Delta(\hat{\beta}_p) = \frac{2\nu_p}{n} - (\frac{\nu_p}{n})^2 - \frac{(\frac{\nu_p}{n})^2}{1 - \frac{\nu_p}{n}} \Delta(\hat{\beta}_p)$

Interpretation $GCV(\hat{\beta}_p) - \sigma^2 - \Delta(\hat{\beta}_p)$ is small if $\nu_p/n$ is small: requires sparsity assumption
Illustration: GCV in LASSO $\nu_\lambda = E(N_1(\lambda))$

\[ \hat{GCV}(\lambda) = \frac{1}{n} \text{SS}_{E}(\hat{\beta}_\lambda) \left( 1 - \frac{\nu_\lambda}{n} \right)^2 \text{ where: } \nu_\lambda = n_1 = E(N_1(\lambda)) \]

\[ GCV(\lambda) = \frac{1}{n} \text{SS}_{E}(\hat{\beta}_\lambda) \left( 1 - \frac{\nu_\lambda}{n} \right)^2 \text{ where: } \nu_\lambda = n_1 = E(N_1(\lambda)) \]

Two results supporting the use of GCV

1. **Uniform efficiency** (requires sparsity)
   
   There exists a sequence $\Lambda_n$ of (large) subsets of $\mathbb{R}$, so that
   
   \[
   \sup_{\lambda \in \Lambda_n} \left| GCV(\lambda) - \sigma^2 - \Delta \lambda \right| / \Delta \lambda + V_n \xrightarrow{p} 0 \text{ as } n \to \infty
   \]
   
   where
   
   \[ V_n = \max \left( 0, \sup_{\lambda \in \Lambda_n} (PE(\hat{\beta}_\lambda, p) - \Delta \lambda) \right) \]

   Interpretation
   
   On a large subset of $\mathbb{R}$, GCV is “equivalent” to Mallows’s Cp (see further)

2. **Behavior near zero $\lambda$ in LASSO (full model)**

   \[
   \lim_{\lambda \to 0} E [GCV(\lambda)] = c_n > 0
   \]

   Interpretation
   
   The behavior near zero does not disturb the minimization procedure

**Uniform efficiency: “equivalence” GCV and Cp**

- **PE (Prediction Error) and Cp**
  
  $E(\Delta_p) = PE(\hat{\beta}_p)$

  (See slide 82)

  Bias-var ⇔ Closeness-complexity

  Equivalence in expected value

- **Cp and GCV**
  
  1. Asymptotic
  2. No expectations
  3. Requires sparsity
  4. Conditions on design matrix $X$

**PART 4: MODEL AVERAGING**